DOES IT MAKE A DIFFERENCE? DATA VISUALIZATIONS AND THE USE OF RESEARCH AND EVALUATION REPORTS.

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATIONAL PSYCHOLOGY

MAY 2018

By

Sena C. P. Sanjines

Dissertation Committee:

Paul R Brandon, Chairperson
Anna Ah Sam
Jack Barile
George M. Harrison
Nicole Lewis

Keywords: Evaluation use, data visualization, symbolic use, cognitive interviews

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members for their thoughtful questions and guidance which helped me create and execute something worth sharing. I would especially like to thank my committee chair, Dr. Paul Brandon, for his excellent support during the writing process.

I would also like to thank the nine individuals who volunteered to participate in interviews and sixteen individuals who volunteered to serve as raters for this study. For a five dollar gift card you gave up hours of your time and made this study possible.

A special thanks goes out to Dr. Stephanie Evergreen who pointed me in the right direction and has served as an invaluable mentor and thought partner over the last three years. I look forward to our continued collaboration.

Last, I would like to thank my two boos. The big one, for taking care of me, our home, and our son when I needed help the most; and the little one, for sleeping through the night so mommy could write.

ABSTRACT

Evaluation use has transfixed the evaluation community since its inception: How do we ensure that the good work we do as evaluators results in something more than a report? Drawing from research on cognition, a new crop of evaluators argued that data visualizations promote the use of evaluations, following the theory that the visuals engage and aid stakeholders in making sense of information (Evergreen, 2011a). This study builds on that theory to explore if the use and quality of data visualizations in research and evaluation reports increases the likelihood that reports will be used. Use, in this case, is an individual referencing a research or evaluation report in legislative testimony on teacher quality. Because use is multifaceted and slippery, I also looked at alternative predictors of use including the length of reports, if they were more like "advocacy" research or "traditional" research, and if the user was affiliated with a university. Using a Poisson regression with frequency of use as the dependent variable, I did not find a relationship between the use or quality of data visualizations and use of reports. However, I did find predictive relationships between the type and length of reports and frequency of use, where longer reports with data visualizations were less likely to be used and reports that were characteristic of advocacy research, (i.e. based on anecdotal evidence, lacking an objective tone, etc.) were more likely to be used.

CONTENTS

TABLES

CHAPTER 1

INTRODUCTION

Five professional standards guide educational evaluation. The first among these is *utility*—the expectation that evaluations should be conducted in a manner that promotes the use of findings (Yarbrough, Shula, Hopson & Caruthers, 2011). Although much of the research on evaluation (RoE) since 1990 has focused on use and the factors that promote it, increasing use of evaluation findings remains a challenge (Alkin & King, 2016; Baughman, Boyd, & Franz, 2012; Brandon & Singh, 2009; Campbell, Townsend, Shaw, Karim, & Markowitz, 2015; Cousins et al., 2015; Fleischer & Christie, 2009; Johnson et al., 2009; Preskill & Caracelli, 1997; Rebora & Turri, 2011; Roseland, Lawrenz, & Thao, 2015; Turnbull, 1999; Weiss, 1998).

In the last few years, some scholars have suggested that the way in which evaluators communicate findings can affect the eventual use of the evaluation, arguing that the use of evaluation findings depends on the degree to which users engage with and make sense of the findings presented (Evergreen, 2011a, Evergreen, 2011b, Evergreen, 2016; Evergreen & Metzner, 2013). Scholars examining evaluation and research communication—the ways in which information from evaluations and research is shared—have found that the way in which humans take in information on a page does not align with the publication norms that social scientists, including evaluators, traditionally follow in reporting (Evergreen, 2011a; Evergreen, 2011b, Kosslyn, Kievit, Russel, & Shephard, 2012; Tufte, 2006). For example, the flagship evaluation journals, the *American Journal of Evaluation* and *New Directions for Evaluation*, require contributors to follow the American Psychological Association (APA) writing guidelines. Among other requirements, the APA (2010) guidelines suggest that authors limit their use of graphs and tables in written publications. In contrast, scholars in cognitive psychology have

found that images on a page or display are better remembered than words (Stenberg, 2006). Though simple, the example illustrates potential disconnects between academic publication norms and how humans process information on a page.

In research, the primary consumers of reports are other scholars familiar with academic writing. In evaluation, the primary consumers of reports are program managers, policymakers, legislators, and other stakeholders who may have limited time or training to make sense of text-heavy, academic writing. In spite of this difference, evaluation reports frequently follow the structure and content of academic publications. For example, a recent evaluation report checklist produced by The Evaluation Center at Western Michigan University mirrors the organization of academic research: table of contents, figures, introduction, background, research design, methods, results, etc. (Robertson & Wingate, 2017). The guide does not mention the design or format of the report itself and the authors refer to *data visualizations* (visual representations of data including graphs) as an alternative form of reporting not covered in the checklist.

In spite of the use of academic publication norms in written evaluation reports, there has been growing interest among evaluators to apply graphic design and data visualizations in reporting. For example, Evergreen (2011a) worked with graphic designers to develop guidelines to incorporate design into evaluation reports. Evergreen and Emery (2014, 2016) extended the work to create design guidelines specifically for data visualizations. Also, the Data Visualization and Reporting Topical Interest Group is one of the largest and fastest growing topical interest groups supported by the American Evaluation Association (AEA).

Although there is growing interest in applying graphic design and using data visualizations in evaluation reporting, research on the effects of these changes has not kept pace. There is limited research on the relationship between how data are presented in reports and how

users intake that information. One exception was a work-in-progress presented at *Evaluation 2016,* the annual conference of the American Evaluation Association (AEA), which explored the relationship between graphic design in reports and use of the reports (Evergreen, 2016). Although Evergreen presented initial findings, the work was incomplete and has not been published. Moreover, Evergreen was interested in the overall design of a report including headings, text, color, organization, etc. The study did not look at the inclusion or quality of data visualizations in reports.

Studying the influence of data visualizations on the use of evaluation reports is an important missing piece for three reasons. First, evaluators appear to be more engaged in the design and use of data visualizations than the design of reports. For example, the majority (79%) of presentations selected for the Data Visualization and Reporting strand at *Evaluation 2017*, the AEA's annual conference, were on data visualization while two (8%) were on the design of reports (American Evaluation Association, 2017). Second, the cognition research upon which scholars have based calls for better design in reporting is grounded in the fact that our eyes are drawn to variations in size, color, orientation, etc. (Evergreen, 2011a; Ware, 2008). In practice, this means that the eyes will be drawn to data visualizations on a page. Finally, while there is limited published research on the impact of data visualizations on use, studies on the interpretation of graphs suggest that the design and quality of data visualizations may affect how data are understood and used (Ali & Peebles, 2012; Ellis and Dix, 2015; Tsuji & Lindgaard, 2014).

Study Purpose

The present study is a response to the increased interest in and use of data visualizations in evaluation reporting, as well as their potential impact on *evaluation communication*—how an evaluation and its findings are shared. I build upon Evergreen's (2016) work-in-progress to explore the relationship between the inclusion and quality of data visualizations in reports and the use of report findings. Drawing from literature on cognition, visual processing, and evaluation use, the underlying theory driving the work is that the use of data visualizations in research and evaluation reports will improve reader engagement and comprehension. This, in turn, will increase the likelihood of use.

The concept of evaluation use is somewhat of a spider web and scholars have been preoccupied with varying types of use since the beginning of evaluation as a profession (Alkin & King, 2017). In the current study, I look specifically at the influence of data visualizations on the *symbolic use* of reports in congressional testimony. Symbolic use refers to the use of an evaluation to persuade or convince others of a stance or opinion (King & Pechman, 1984; Johnson et al., 2009) or justify action or inaction (Henry & Mark, 2003). I chose to focus on the symbolic use of research and evaluation reports due to the nature of the existing data I used in the study. Moreover, reports referenced in congressional testimony have the potential to impact national policy and factors influencing their use deserve attention.

I address two research questions exploring (a) if the use of data visualizations in reports increases the likelihood reports will be used, and (b) if the quality of data visualizations in reports increases the likelihood reports will be used. I do this by using the Data Visualization Checklist, or DVC (Evergreen & Emery, 2016), to rate reports referenced in congressional testimony where a report is considered *used* if it was referenced (See Chapter 3 for a full

description of the reports used for the study). Some reports were referenced more than others. Following this, I use the frequency with which reports were referenced as the dependent variable. In addition to DVC ratings, I explore alternative predictors to help explain use including (a) the length of the report, (b) if the report is classified as an advocacy or a traditional research report, and (c) the academic affiliation of the user. Because this is the first time the DVC is used to measure the quality of data visualizations in a study, I also collect evidence that the DVC is a measure of the quality of data visualizations. I do this by conducting cognitive interviews with individuals likely to use the checklist and document the extent to which their understanding and use of the tool aligns with its intent.

This study contributes to the growing pool of research on evaluation. It is the first empirical study to explore the potential impact of using data visualizations in evaluation reports on evaluation use. Also, the study will contribute to our understanding of the symbolic use of evaluations. In an essay on the future of research on evaluation (RoE), Azzam and Jacobson (2015) called for increased research on evaluation communication and an overall call for improved documentation of core evaluation functions, including communication. The results of the study will answer their call and add to what we know about evaluation communication and evaluation use. In addition, the work provides evidence about the validity of inferences from the DVC for assessing data visualizations in research and evaluation reports, which may aid future research on evaluation communication.

Definition of Terms

Report. A report is a written document providing data or information about a specific topic or study. In the present study, *report* refers to evaluation and social science research reports on the topic of teacher quality referenced in congressional testimony. I combine evaluation and

research reports due to the overlap in content and form (Evergreen, 2011a, Evergreen 2011b) and because I am exploring symbolic use of reports in congressional testimony, an arena that does not distinguish between research and evaluation.

Data Visualization. Data visualizations are defined as images based on data or statistics that are representative of the data, readable, and support the exploration, examination, and communication of the data (Azzam, Evergreen, Germuth, & Kistler, 2013 p. 9). As such, any graphics or images that do not represent data, for example icons, are not considered data visualizations.

CHAPTER 2

LITERATURE REVIEW

The core of my study is research on evaluation (RoE), specifically looking at evaluation use. In this chapter I situate my study as RoE and review the major literature on evaluation use. I then discuss research on evaluation communication, as well as key concepts of cognition, graphic design, and data visualization that I drew from in the design of the study. I end the chapter with a review of my search for existing research on data visualization and evaluation use before discussing how this study will contribute to existing literature on evaluation use and RoE.

I used slightly different approaches when searching for literature on evaluation use and evaluation communication, described in each section. I also describe the foundational literature from research on cognition and graphic design. However, I did not conduct a systematic review of the literature in this area for two reasons. First, others have already conducted thorough reviews of the literature on cognition and design in relation to data visualizations (See Evergreen, 2011a; Ware 2008, 2012). Second, my research question is broadly focused on the relationship between data visualizations and report use, not on specific aspects of graphic design. Therefore conducting a thorough analysis of the research on cognition is outside the focus of this study.

Research on Evaluation

Fifteen years ago, Henry and Mark (2003) bemoaned what they perceived to be a diminished practice of systematic data collection and assumption testing within the evaluation field. However in recent years there has been an increase in published RoE studies (Coryn et al., 2017; Vallin, Philipoff, Pierce, & Brandon, 2015), as well as an increased demand for RoE

(Lewis, Harrison, Ah Sam, & Brandon, 2015). For example, Vallin, Philippoff, Pierce, and Brandon (2015) reviewed work published in the *American Journal of Evaluation* (AJE) between 1998 and 2014 and found the percentage of articles classified as RoE increased through 2011. The authors found that most RoE published in the flagship journal was descriptive in nature, consisting of single case studies or exemplars. In a broader and more restrictive review of RoE, Coryn et al. (2017) reviewed empirical research published in 14 evaluation-focused journals between 2005 and 2014 and also found that most RoE was descriptive. Thus, while there appears to be an increase in published RoE, the studies are largely descriptive and therefore have limited generalizability beyond informing best practice.

More recently, Azzam and Jacobson (2015) echoed Henry and Mark (2003) and called on evaluators to embed systematic data collection and documentation of major evaluation activities into their evaluation work as a necessary next step in RoE. Azzam and Jacobson also pointed out that common data on evaluation use should be regularly collected across evaluations including, but not limited to, data on evaluation communication. Nearly a decade earlier, Johnson et al. (2009) also called out the importance of research on evaluation communication. Though both Johnson et al. and Azzam and Jacobson (2015) were concerned with overall communication between evaluators and the evaluand, Azzam and Jacobson specifically pointed out the potential benefit of methods or frameworks to define and measure effective evaluation communication (p. 109).

Research on Evaluation Use

The present study is situated as research on evaluation use. However, this is far from the first study interested in evaluation use. In the introduction to Alkin and King's (2016) review of evaluation use, the authors reminded readers that the history of evaluation parallels that of

human history and that we have been interested in the use of evaluation for nearly as long. Due to the prevalence of research on evaluation use, a number of scholars have published reviews of the research. As such, I restricted my review of literature to published reviews of research or literature on evaluation use.

I identified reviews of literature on evaluation use in a three-step process. First, I searched in major evaluation journals including *AJE, New Directions for Evaluation*, *Evaluation Review*, and *Evaluation and the Health Professions* for articles using the terms "evaluation" and "use", "utilization", or "influence" in the work's title. I next extended the search to Academic Search Complete database for articles published outside the field of evaluation.[1] I then reviewed titles and abstracts to select articles that were reviews of evaluation use literature or research. Last, I searched the bibliographies of the selected articles for any additional reviews not identified in the journal or database search.[2] Below, I discuss common threads I identified across the nine reviews of research or literature on evaluation use. The selected articles, description of the work, definitions of use included in the review, and factors related to use included in the review are presented in Table 2.1.

Definitions of Evaluation Use

To date, there is no universal definition of evaluation use (Alkin & King, 2016). However, most scholars have defined use in terms of its function, often referred to as type of use. Over the years, three primary types of use—*instrumental*, *conceptual*, and *persuasive*—have

[1] A general search of the term "evaluation use" in Academic Search Complete resulted in over 4,000 entries of limited relevancy. I then added the search term "program evaluation" as subject term which narrowed the results to 14 records.

[2] Two reviews of research on evaluation use Brandon & Singh (2009) included in their study—Thompson and King (1981) and Cousins (2003)—were not identified in my search. I did not include these two studies in my review because Thompson & King's work was an unpublished conference paper and Cousin's (2003) work was a review of research on participatory evaluation and, as such, did not fit the search criteria.

been consistently defined and discussed (Alkin & King, 2017; Cousins & Leithwood, 1986; Johnson et al., 2009; Leviton and Hughes, 1981; Shulha & Cousins, 1997; Weiss, 1998). Importantly, scholars have agreed that factors contributing to use are somewhat dependent on the type of use under investigation (Leviton & Hughes, 1981; Cousins & Leithwood, 1986). Following this, I further explore major types of use in the *Types of Use* section below.

Expanding definitions of use. While there is no one widely accepted way to define evaluation use, there has been a pattern over time that definitions must shift and expand in order to capture changing notions of what counts as use. For example, early writings defined use in terms of the direct use of evaluation findings to either make decisions, increase understanding, or to persuade others for some action or inaction (Cousins & Leithwood, 1986; Leviton & Hughes, 1981). Later reviews expanded our understanding of evaluation use to include changes occurring as a result of participating in an evaluation, known as *process use* (Shulha & Cousins, 1997; Johnson et al., 2009). In addition, scholars attempting to account for the complexity within evaluation use introduced the term *evaluation influence* to push beyond limitations of the term "use" (Kirkhart, 2000), which we see addressed as early as Johnson et al.'s (2009) review.

Factors Influencing Evaluation Use

Most literature reviews of evaluation use have included factors found to contribute to use, as summarized in Table 2.1 (Alkin & King, 2017; Cousins & Leithwood, 1986; Johnson et al., 2009; Leviton and Hughes, 1981; Shulha & Cousins, 1997). Across reviews, there has been considerable overlap in the factors believed to contribute to use (Alkin & King, 2017), including but not limited to the quality of the evaluation, relevance to stakeholders, communication between the evaluator and users, and contextual factors including users commitment to the evaluation. Of note, only the reviews by Cousins and Leithwood (1986) and Johnson et al.

(2009) employed rigorous systematic reviews of the literature. Johnson et al. built upon the work of Cousins and Leithwood and used the same framework to review studies. For both, due to the variability in definitions and measures of use, metaanalyses of the research were not possible. However, both studies identified a handful of factors that were well-supported in the research and appeared to have strong relationships with use. These included the quality of the evaluation, the decision-making setting, users' commitment to the evaluation, and the relevancy of the evaluation to users. Reflecting the expanded notion of evaluation use that occurred in the time between the two reviews, Johnson et al. also found that stakeholder involvement was a key factor that increased the capacity of those involved to use the evaluation. Moreover, Johnson et al. specifically argued evaluators should engage, interact with, and communicate with stakeholders to increase the likelihood of use.

Table 2.1

Reviews of Research and Literature on Evaluation Use

Study	Description	Definition(s) of use	Factors affecting use
Leviton & Hughes (1981)	Summarized findings from undefined (no years/sector noted) evaluation and social science research.	Use of evaluation results for program or policy. Three types: instrumental, conceptual, and persuasive.	• relevance • communication • information processing • credibility • user involvement/ advocacy.
Cousins & Leithwood (1986)	Reviewed 65 studies on evaluation use from education, health, and social services sectors published between 1971 and 1985.	Evaluation use defined in terms of type of use: decision-making, education, or mental processing (p. 332). Potential for use was added after reviewing the research.	Created a framework of 12 factors related to use, 6 associated with the evaluation, and 6 with the context of use. In particular, they found relevance of the evaluation and data collected, alignment with expectations,

Study	Description	Definition(s) of use	Factors affecting use
			involving users, and having minimal conflicting data, or credibility, aided use.
Shulha & Cousins (1997)	Synthesized the main themes within research on evaluation use between 1986-1996.	Use is multidimensional and largely consisting of instrumental (decisions), conceptual (education), and symbolic (political) functions.	Focused on themes within the research on evaluation use: • importance of context; • consideration of process use; • collaborative approaches to evaluation; and • framing misuse.
Weiss (1998)	Summarized the state of evaluation use. Note, Weiss' oft-cited review included few study citations and instead painted broad strokes about the field.	Evaluation use defined in terms of type of use: instrumental (decision-making), conceptual (new understandings), persuasive (support a position), and use outside the program under study.	Did not consider factors of use. Focused on defining types of use, elements used (findings, process, etc.), and users to include both individuals and organizations.
Caracelli (2000)	Review of historical markers in the evaluation field over a 30-year period in order to situate other chapters in a NDE volume on expanding the scope of evaluation use.	Not provided, but she argues that conceptions of use must be broad enough to include multiple perspectives of use – considering use and influence together.	Did not consider factors of use. Focused on history of evaluation to understand expanding definitions of its use.
Brandon & Singh (2009)	Reviewed methodological soundness of studies cited in five reviews of evaluation use.	Not provided.	Did not consider factors of use. Found weaknesses in the content-related validity of reviewed studies and in the balance of types of methods for studying use.
Johnson et al. (2009)	Reviewed research on evaluation use published between	Defined use as "the application of evaluation processes,	Identified stakeholder involvement as an emergent factor, and

Study	Description	Definition(s) of use	Factors affecting use
	1986 and 2005 using Cousin & Leithwood's (1986) framework. The authors' inclusion criteria included study quality.	products, or findings to produce an effect" (p.378). The authors considered use in terms of what was used—that is, process or findings use.	one that had a mitigating effect on other influences on use.
Herbert (2014)	Review of 28 empirical studies that used the concept of evaluation influence to frame the study.	Presented multiple definitions of evaluation influence drawn from Kirkhart (2000), Henry and Mark (2003), and Mark and Henry (2004)	Did not consider factors of use. The review was concerned with the viability of evaluation influence as a conceptual basis for study.
Alkin & King (2017)	Summarized the development of evaluation use over a 40-year period. The article is the second in the three-part series and includes factors affecting use.	Rather than a singular definition, the authors provide a sentence map including what is used (findings or process), users, how it is used and for what purpose in order to define use. They gently reject the argument for use of the term evaluation influence.	Identified four groups of factors: • User, • Evaluator, • Evaluation, and • Organizational/ social context factors

Types of Evaluation Use

The earliest writings on evaluation use by Carol Weiss in the 1960s and 1970s came from a social science research stance and defined evaluation use as a program's use of findings and/or recommendations produced by an evaluation (Weiss, 1967, 1972, 1979). In other words, *use* was understood exclusively as the application of evaluation findings to make decisions and take action within the program under study. Coming from a different perspective, Marvin Alkin identified use as core to evaluation and that which distinguished the field from traditional research (Alkin, Daillak, & White, 1979; Alkin & King, 2016). In spite of different perspectives

on the concept, both scholars were uniform in initially defining evaluation use as the use of information for program decision-making (Alkin et al., 1979; Weiss, 1967, 1972), before later expanding their definitions to consider use beyond direct program-level decisions (Alkin & King, 2016; Weiss, 1998). Both Weiss and Alkin approached the concept of use in terms of the who and what: Who is using information garnered from an evaluation, and for what purpose (Weiss, 1998; Alkin & King, 2017)?

Describing evaluation use in terms of who is using evaluations and for what purpose, the literature naturally focused on how an evaluation is used in order to define what is meant by it. Below I outline the primary types of evaluation use. An overview of these, including examples of each, are presented in Table 2.2. I drew the definitions and descriptions outlined below and presented in Table 2.2 from Leviton & Hughes (1981), Weiss (1998), and Shulha and Cousins (1997).

Instrumental use. Instrumental use is often what comes to mind when thinking about evaluation use. It is the direct application of evaluation findings to make decisions about a program. It is also the most studied (Johnson et al., 2009). In early writings about evaluation use, scholars have sometimes discussed the difference between instrumental and conceptual use in terms of being able to document use (instrumental) or not (conceptual).

Conceptual use. Conceptual use is just that, conceptual – or in the head. Scholars describe conceptual use as a change in knowledge or understanding as a result of an evaluation. Most discuss this as education resulting from evaluation findings, though Weiss (1998) also included understandings garnered from participation in an evaluation in her definition of conceptual use. However, the consensus among scholars is that conceptual use refers to a change in knowledge or understanding based on the findings of an evaluation.

Persuasive and symbolic use. Persuasive use refers to the use of an evaluation to justify a position or action/inaction (Leviton & Hughes, 1981). What frequently comes to mind is a director or policymaker who already has a plan of action a priori and uses evaluation findings to justify it. However, in one of the earliest definitions of the term, Leviton & Hughes (1981) promoted the use of evaluations to justify decisions and argued that reference to an evaluation without discussion or dialogue— political maneuvering or paying lip service to the findings— violates the definition of use.

Symbolic use is closely related to persuasive use, and scholars have used the terms interchangeably. For example, Shulha and Cousins (1997) presented the first review of evaluation use to include the term *symbolic use*, which they defined by referring readers to Leviton and Hughes' (1981) description of persuasive use. Overall, scholars have used the term *symbolic use* to classify the use of an evaluation to persuade or convince others of a stance or opinion (King & Pechman, 1984; Johnson et al., 2009), or justify action or inaction (Henry & Mark, 2003). Alkin and Taut (2003) argued that the difference between persuasive, which they referred to as *legitimative,* and symbolic use is that persuasive use is the use of findings, and symbolic use is the use of the evaluation processes, or the fact an evaluation is occurring.[3] Regardless of this distinction, scholars have used the terms interchangeably, with more recent literature favoring the term *symbolic* (Alkin & King, 2016).

Process use. Process use refers to new understandings about the process of evaluation garnered by those involved in the evaluation as a result of involvement in evaluation activities. The concept came largely from the work of Michael Quinn Patton (1997), who observed

[3] Of note, simply referring to the fact that an evaluation is taking place in order to justify a position or action may be considered paying lip-service to the evaluation which Leviton and Hughes (1981) explicitly argued would not be considered use of an evaluation.

increased evaluative capacity of individuals involved in an evaluation. The term has since been used more generically to refer to use that occurs as a result of participating in or conducting an evaluation (Alkin & King, 2016).

Table 2.2

Types of Evaluation Use

Type of use	**How the evaluation is used**	**Who is the user**	**Example**
Instrumental	Information from an evaluation, often evaluation findings, used by individuals involved with a program or policy to make decisions that directly affect the program.	Individuals in decision-making roles, usually involved with the program.	A program director used the results of an evaluation to adjust the training protocol of staff to improve the consistency of program interventions.
Conceptual	Information from an evaluation causes a change in knowledge or understanding about the program or policy.	Individuals in decision-making roles or others involved in the program, (program stakeholders).	A teacher who participated in an evaluation about an after-school program developed an understanding of what the program was trying to achieve as a result of the evaluation.
Persuasive or Symbolic	Information from an evaluation or the act of completing an evaluation is used to legitimate or garner support for a position or decision.	Often individuals in decision-making roles.	A director wants to make a change in program direction and uses the results of an evaluation to persuade her board and her staff that it is the right decision.
Process	Participating in evaluation activities causes a change in knowledge or understanding and/or	Individuals involved in the program and/or the evaluation, (program stakeholders).	A group of teachers change the way they recruit students into a program after identifying it as a

Type of use	How the evaluation is used	Who is the user	Example
	builds evaluation capacity.		problem during a focus group discussion.
Use beyond the bounds of a program	Evaluation results, combined with results from similar evaluations, are used to influence public policy or shift existing schools of thought.	Individuals, usually in influential roles such as policy-makers, thought leaders, and so forth but not directly involved with an evaluation.	A government official cites similar results from a number of evaluations to persuade Congress to increase funding for early childhood education.

Use beyond the bounds of a program. Weiss' oft-cited (1998) review of evaluation use included an additional type: the use of an evaluation beyond the bounds of a program to influence policy change. She specifically described this type of use as a ground swell of similar work that could help sway or influence policy or programs outside of the evaluation. Although this distinction was not picked up by other scholars, I bring it up here to make the point that types of use have been defined by their functions, (decision-making, education, persuasion, etc.) rather than their place of use (inside or outside of program).

Misuse. Misuse has been discussed in lock-step with use since some of the first publications on evaluation use (Leighton & Hughes, 1981). The theme across references to misuse is that there are different kinds of misuse, with some more malicious than others (Shulha & Cousins, 1997), and all of which must be attended to. Alkin and King (2017) argued that use and misuse are separate concepts rather than a continuum and, as such, should be considered and researched separately. For this reason, I did not include misuse in Table 2.2.

The Reality Check

While I present types of evaluation use in neat little boxes, in reality use is complicated and multifaceted. There are multiple types of use by various individuals and/or groups of users occurring concurrently during and following an evaluation (Shula & Cousins, 1997; Alkin & King, 2016). In a review of empirical research on evaluation use focusing only on instrumental use of evaluation outcomes, Cousins and Leithwood (1986) found a wide variety in concepts of use and users across the 65 studies reviewed. King and Pechman (1984) summarized the disconnect between assumptions of use and actual evaluation use well: "Evaluation use may appeal more to a sense of the way things *should* be than to an awareness of how things *are*" They added the "big bang myth" of evaluation use is that results will lead to immediate and observable change (p. 242, emphasis in original).

The sticky spider web of use in practice drew scholars to question if the term "use" was the best way to represent the effects of an evaluation on the evaluand, those involved, or others beyond the scope of an evaluation. Kirkhart (2000) argued for a change in language to discuss use as *influence* in order to better capture the nuanced impact and wide reach of evaluation practice and findings. Henry and Mark (2003) also promoted the term influence to better capture the variation in how evaluations may impact individuals, organizations, and society while avoiding the need to continually identify and define new types of use. Considered together, the primary critique offered by proponents of evaluation influence is that the terms use and utilization are loaded and far too narrow to constitute the potential impacts of an evaluation. While definitions of evaluation use overtime have focused on users and what the evaluation is used for, proponents of evaluation influence have purposefully avoided defining the parameters of influence (Johnson et al, 2009). Following this, the consensus across reviews of research that

included evaluation influence was that the concept was difficult to define and research (Alkin & King, 2017; Johnson et al., 2009; Herbert, 2014).

Drawing from definitions of evaluation use and influence, I use the term *symbolic use* in my study and define it as the persuasive use of an evaluation, whether the results or the act of evaluation itself, by program decision makers or individuals outside of a program to justify a position or action.

Evaluation Communication

Recently, scholars have begun to argue that the way in which evaluation findings are presented may affect use of an evaluation (Azzam, Evergreen, Germuth, & Kristler, 2013; Evergreen 2011a, 2011b, Evergreen & Metzer, 2013). Interest in the design of evaluation reports and presentation of findings is not new (Newman, Brown, & Braskcamp, 1980; Torres, Preskill, & Pontiek, 1996). However, in spite of the long history of interest in the area, existing research on evaluation communication is largely descriptive (Azzam & Jacobson, 2015; Evergreen, 2011a; 2011b). Prior studies focused on the cognitive challenges of existing report platforms, including written reports (Evergreen, 2011a) and electronic slide shows (Kosslyn, et al., 2012; Tufte, 2006). For example, Kosslyn et al. (2012) reviewed a random selection of slide show presentations from academic research, education, government, and business and found that, on average, presentations violated six cognitive communication principles based on processes of encoding, working memory, and accessing long-term memory. In a more extreme example, Tufte (2006) argued communication challenges due to NASA's use of PowerPoint led to the Challenger tragedy.

Evergreen (2011a) rated the use of graphic design principles (headings, text layout, etc.) in a random sample of evaluation reports to determine the extent to which the written reports followed cognitive communication principles. She found that the way information was presented in the reports actively worked against reader comprehension and argued that the report authors missed opportunities to fully engage their readers. She referred to this as the "communication-cognition gap" and pushed for evaluators to follow the lead of graphic designers to improve evaluation communication (p.2).

Data Visualization and Cognition

Graphic design principles—for example the use of color, shapes, line, formatting, and so forth—are most applicable in the use and design of graphs, charts, and images in research and evaluation reports, commonly referred to as data visualizations (Evergreen, 2011a). Scholars promoting the use of design in evaluation communication consistently draw on research from cognitive science related to attention and working memory to justify the call for improved data visualizations (Evergreen, 2011a; Evergreen, 2011b; Evergreen & Metzer, 2015). As mentioned at the beginning of the chapter, I am not summarizing reviews of the research on cognition and graphic design that have been produced elsewhere (see Ware's 2008 and 2012 books for thorough reviews). Instead, I present an overview of key concepts from cognition literature that are important to the underlying theory of my study including; noticing, the role of chunking information, and cognitive load.

How Visualized Data Aids Cognition

What we know about visual processing and working memory indicates data visualizations may aid in users' ability to notice and make sense of data. Our eyes are constantly

scanning our environment and are very efficient at noticing variations or disruptions in our field of vision, discussed as *active vision* or *preattention* (Evergreen, 2011a; Ware, 2008, 2012). In research and evaluation reports, it is this function of vision that causes our eyes to be drawn to data visualizations or headings when scanning a page (Evergreen, 2011a).

Beyond catching our attention, research on cognition has found that the use of visual images aid in the processes of encoding information and mental recall (Radvansky & Ashcraft, 2016; Sternberg, 2006; Ware, 2008, 2012). The primary way in which data visualizations support these functions is by chunking multiple pieces of information that may reduce cognitive load on working memory during the encoding process (Ware, 2008). I say "may" because the complexity of the chunks or objects makes a difference in how much information may be retained (Xu and Chun, 2006).

Images and memory. It is widely understood that images are better remembered than text (Sternberg, 2006). In his work, Sternberg (2006) pointed out that research supporting the superiority of images over text for mental recall has been around since the late 1800s. However, in order to aid in recall, information in data visualizations must be encoded into long-term memory (Radvansky & Ashcraft, 2016).

Miller (1956) famously argued that humans have a capacity to retain seven chunks of information at any given time, plus or minus two. Sperling's (1960) work on the recall of visual information found that the number dropped to about four, which was also supported by Xu and Chun (2006). Ware (2008) went even farther and argued visual working memory can hold approximately one to three objects simultaneously. The consensus is that, while imagery helps the human mind encode new information, processing visual information takes more mental load. For example, Chandler and Sweller (1991) explored the concept of cognitive load through a

series of experiments pairing diagrams with explanatory text and found extraneous information included with images reduced understanding.

Overall, scholars have found that while images are remembered better than text (Sperling, 2006; Ware, 2008, 2012), processing images takes more mental energy (Sperling, 1960; Xu & Chun, 2006). For these reasons, how information is presented, or the design of data visualizations, may impact the uptake and processing of information.

The Importance of Design

Ware (2012) argued there are two types of representations affecting visual processing; sensory and arbitrary. Sensory representations are those tied to the inherent ways our brains take in information—that is, variations in color, shape, size, etc. Arbitrary representations are based on learned information and are often context-bound. While data visualizations employ both types of representations, Ware argued that well-designed data visualizations manipulate sensory representations such as color, shape, etc. to aid our brains in processing information. In addition to design elements, Chen and Yu (2000) conducted a meta-analysis of 35 empirical studies on information visualization and found that individuals with similar cognitive abilities performed better (according to accuracy or efficiency) with simpler visual-spatial designs. While Chen and Yu's study was laboratory-based and used visualizations generated from information design systems, their findings aligned with arguments from evaluation scholars that simple designs reduce cognitive load and aid in comprehension (Evergreen, 2011a, 2011b; Evergreen & Metzer, 2016).

In a direct connection between working memory and interpretation, Halford, Cowan, and Andrews (2007) argued that limitations to working memory, —the ability to hold multiple pieces

of information at once—are also true for reasoning. The argument was based on both functions being dependent on an individual's ability to "form and preserve bindings between different pieces of information" (p. 236). While research continues, their findings suggest that visualization decisions which reduce cognitive load on working memory may similarly aid in reasoning by reducing the amount of information individuals must hold and connect during the sense-making process.

Misinterpretation of Data Visualizations

Design decisions may not only affect the ease of interpretation. They may also affect readers' ability to correctly interpret the data presented. Drawing from theories on cognition and data visualization, Ali and Peebles (2012) conducted a series of experiments looking at students' comprehension of bar graphs and line graphs and found that students were significantly more likely to misinterpret data in line graphs. Importantly, Ali and Peebles interpreted their results according to Gestalt principles—how individual features of graphs are grouped together in visual processing to make a coherent whole. After revising the graphs to better align with Gestalt principles, they found a significant improvement in students' performance. Their findings suggested that while the type of graph used may impact accuracy of interpretation, changes in design can also correct for the problem.

In addition to complexity and graph type affecting interpretation of data visualizations, Tsuji & Lindgaard (2014) found that users' level of experience also played a part. They compared the ability of novices (undergraduate students) and experts (PhD students) in business and psychology to explain graphs from their respective fields and found an expertise effect on the time it took for participants to explain the graphs. They also found that experts were able to more completely explain the information in the graphs than novices.

Adding to the conversation, Ellis and Dix (2015) explored the effect of uncertainty on decision-making, where the uncertainty was due to difficulty in comprehending graphs. They argued that when the data in a graph are not clear at a glance, users must resort to spatial processing and working memory to make sense of the information. This, in turn, takes more cognitive load and increases the likelihood of misinterpretation of the data due to cognitive bias—for example seeing a pattern where there is none or sticking with a value or judgment that is top of mind. While Ellis and Dix were predominately interested in visualizations for data analysis, their arguments are particularly relevant for program evaluation in which users are often decision-makers who may experience uncertainty in reading data in graphs.

Taken together, theories and research on information communication indicate multiple factors can lead to user misinterpretation of data visualizations, including complexity, graph type, user's level of experience, and just the act of having to connect the dots in the data.

Design of Data Visualizations

Evaluation scholars have promoted the use of specific design elements in data visualizations to aid in reader comprehension. Table 2.3 provides a summary of data visualization recommendations from recent scholarship. At a glance, we see alignment in the elements considered important to aid in reader comprehension including text, arrangement, color, and lines, or ink on a page. Specifically, evaluation scholars promote the use of text to aid interpretation, the arrangement of elements into meaningful groups or order, the use of color to draw attention, and removal of unnecessary lines or ink that may compete for a reader's attention.

Table 2.3

Design Elements to Aid in User Comprehension of Data Visualizations

Element	**Evergreen & Metzer (2013)**	**Evergreen & Emery (2016)**	**Pankaj & Emery (2016)**
Text	Use design, including text, to guide what an individual should notice in a visual display.	Text that is used must be clear, concise, and include the takeaway message.	Use text for labeling and titles and subtitles. Titles should be generic for analysis and interpretive in a final report.
Arrangement		Visualization elements should follow a thoughtful arrangement, for example. sorted in a meaningful order	Group data visualizations by common themes or topic to aid in analysis.
Color	Use design, including color, to guide what an individual should notice in a visual display.	Color should be used to highlight patterns or guide a reader's eyes to key parts of the display.	Use color to draw attention to or de-emphasize elements in a visualization. Coloring for analysis should not emphasize any one pattern while coloring in a final report should.
Lines/ Simplification	Strip away any information that is not essential for reader understanding.	Gridlines, borders, tick marks, etc., add clutter to a display and should only be used if they add needed information.	

Of the three articles referenced, only Evergreen and Metzner (2013) directly connected their design recommendations to findings from research on cognition and information processing (Chen & Yu, 2006; Cowan, 2001; Sternburg, 2006; Ware, 2012; Xu & Chun, 2006). Evergreen and Emery's (2016) and Pankaj and Emery's (2016) recommendations were based on a combination of uncited research and personal experience. However, the recommendations based

on research and those based on practice are closely aligned and include elements identified important to cognition such as color and ink on the page (Ware, 2008; Chen & Yu, 2006). Considered holistically, decisions in the design of data visualizations should largely be made to (a) draw attention and (b) reduce "noise" where noise is anything extraneous.

Data Visualization and Evaluation Use

I searched for existing empirical studies on the relationship between data visualizations and evaluation use or use of research in EBSCO databases including ERIC and Psychological and Behavior Sciences, as well as flagship evaluation journals *American Journal of Evaluation* and *New Directions for Evaluation,* using the search terms "data visualization" and "graphic design." Through the search, I found scholars have explored the use of data visualizations to aid in pedagogy (Ealy, 2016), data analysis (Erwin, Bond, & Jain, 2015), and the study of teaching graphic design as a discipline (Powell, 2013). After extending the search to Academic Search Premier to include biomedical journals, I found additional studies that looked at the relationship between data visualizations and user comprehension (Skau, Harrison, & Kosara, 2015) and engagement (Obe, 2013). However, I did not find any published empirical studies that explored the use of data visualizations in social science research or evaluation in relation to use of research or evaluation findings.

At the American Evaluation Association annual conference, *Evaluation 2016,* Evergreen (2016) presented a work-in-progress in which she explored the relationship between overall report design (headings, layout, etc.) and use of report findings. Evergreen's work appears to be the first to explore the relationship between the use of graphic design in research and evaluation reports and use of report findings. While she addressed overall design including type,

arrangement of text, use of color, and use of graphics, she did not examine data visualizations in the reports.

In addition to Evergreen's (2016) work-in-progress, there is an entire literature strand in health sciences devoted to the concept of *knowledge translation*, the term given to applying research to practice. In a review of the intersection between health researchers and policy-makers, Marten and Roos (2005) found graphical representation of data, discussed as "evidence-based storytelling," supported use of health sciences research (p. 78).

Summary

Research on evaluation has been preoccupied with understanding if and when evaluations are used and the factors that promote use (Alkin & King, 2016). However, most of this research has focused on instrumental use, or the direct use of findings to inform decisions about a program (Cousins & Leithwood, 1986; Brandon & Singh, 2009; Johnson et al., 2009), and process use, or new understandings resulting from participation in an evaluation (Alkin & King, 2016; Johnson, et al., 2009; Shulha & Cousins, 1997). What we know from prior research is that stakeholder engagement throughout an evaluation supports the use of evaluation findings (Johnson et al., 2009; Shulha & Cousins, 1997).

What we do not yet fully understand are the factors that promote symbolic use of evaluations by individuals outside of a program including policymakers, legislators, and other stakeholders. Based on prior reviews of evaluation use, I defined symbolic use the persuasive use of an evaluation, whether the results or the act of evaluation itself, to justify a position or action. In situations where the user of an evaluation is not involved in the program evaluated,

information is largely shared through written reports which often mirror the content and conventions of academic publications (Evergreen, 2011a).

Though limited, prior research on evaluation communication suggests that the way in which information is presented affects reader's intake of the information (Evergreen, 2011a; Kosslyn et al., 2012; Newman, Brown, & Braskamp, 1980; Tufte, 2006). Also, what we know about visual processing and cognition suggests that data visualizations may help readers notice, make sense of, and remember information presented in reports (Ware, 2008, 2012). However, the design of data visualizations matter and complex visualizations are likely to impede understanding (Chen & Yu, 2000). Thus, the underlying theory driving my study is that good data visualizations in reports will improve reader engagement with and comprehension of the information presented, which in turn will increase use of that information. I explore this theory by investigating if the use and quality of data visualizations in reports referenced in congressional testimony was related to the frequency the reports were referenced.

CHAPTER 3

METHODS

I posed two research questions to explore my theory that data visualizations are related to the use of reports. They are:

1. Does the *use* of data visualizations in research and evaluation reports increase the likelihood that report findings will be used; and
2. Does the *quality* of data visualizations in research and evaluation reports increase the likelihood that report findings will be used?

I conducted a rating study to address the research questions and explore relationships between the use and quality of data visualizations and use of reports. In order to assess the quality of data visualizations in reports, I used the Data Visualization Checklist (DVC; Evergreen & Emery, 2016), discussed further below. This was the first time the DVC was used for research. As such, I also collected evidence of construct validity and reliability of the tool.

As mentioned in Chapter 2, use is a sticky spider web and there are many potential factors contributing to the likelihood that an evaluation will be used. For this reason, I included additional variables related to characteristics of the reports and the report users as alternative explanations of use (Alkin & King, 2017; Newman, Brown, & Braskamp, 1980). These included the type of report, user affiliation, and length of reports.

In this chapter, I describe the data used in the study, the rationale for and description of the alternative predictors of use, and the analyses I used to answer the research questions. I conclude the chapter with a description of the methods I used to collect validity evidence—as well as the results—for the DVC as a measure of data visualization quality.

Report Data

The primary challenge in tracing use of research and evaluation reports is defining what counts as use. Decisions usually happen over time and are influenced by a number of factors. The added challenge of investigating use of an evaluation outside the bounds of a program is tracking use, akin to following the path of research post-publication. Cousins et al. (2015) called tracing use of academic research a "notoriously difficult task" (p. 75). In order to address this challenge, I used existing data from research conducted by Reckhow, Holden, and Tompkins-Stange (2015), which included a clearly defined and documented "use" variable[4].

Reckhow et al. (2015) examined the influence of think tanks and advocacy research on education policy, specifically policies on teacher quality. Among their research questions was how often advocacy research was referenced in congressional testimony. They addressed this question by reviewing publically available congressional testimony related to teacher quality from 2000 to 2015. Through this process they identified 600 reports submitted as testimony. These constitute the full population of reports used for my study.

Although all of the 600 reports identified in the study were referenced, some reports were referenced only once while others were referenced dozens of times. Because of this variation, I used frequency of use as the dependent variable for the study, discussed further in the analysis section.

Sample

The original dataset from Reckhow et al. (2015) included 600 reports referenced in legislative testimony. Evergreen (2016) drew from this dataset for her preliminary study and

[4] I received approval to use data for my study from Dr. Reckhow via email on September 1, 2016.

rated the overall design of 89 reports. Because the proportion of reports referenced more than once was very small in the original dataset, Evergreen used a modified random sampling frame and included all reports that were referenced more than once plus a random sample of those reports referenced only once.

Prior to sampling for my study, I standardized report titles by removing quotation marks and separated the report title and report producer. Through this process, I identified duplicate reports not previously identified, potentially due to the combination of (a) report title and author in a single field; (b) variations in capitalization; and (c) variations in full and shortened report titles. I also identified citations that were edited books. For the purposes of this study, a book was defined as a volume over 300 pages separated into chapters or a volume categorized by book sellers, libraries, etc., as a "book." Based on duplicate entries I created a new frequency of use variable accounting for duplicate records not previously identified. For example, if the frequency of use for a given report was "2," in other words referenced twice, and there was a duplicate report with a frequency of use as "1," I removed the duplicate entry and revised the frequency of use for the original entry to "3." After removing duplicates and books identified prior to sampling, there was a total of 562 reports in the full dataset.

I used a two-step process to select reports for the analytic sample. First, I included the reports Evergreen (2016) used in her study. Her sample included all of the reports referenced more than once plus a random sample of 63 reports referenced one time.[5] Second, I included any reports referenced more than once not originally included in Evergreen's sample due to the duplicate records that had not been identified previously. I then employed Evergreen's sampling

[5] Seven of the reports Evergreen (2016) included in her study were duplicates, and as such, were incorrectly classified as reports referenced once.

frame to increase the number of randomly sampled reports referenced one time to ensure a 95% confidence level in the sample. The resulting analytic sample included 278 reports—59 reports (21%) referenced more than once and a random sample of 219 reports (79%) referenced one time drawn from an overall population of 504 reports referenced once.

Analytic sample. I obtained PDF copies of reports for analysis through online searches of report titles using Google Search as well as via periodical and organizational websites. I used publication information available in the source file, including the report's author, publisher, and date published to verify the PDF copies identified through the online search were the correct documents. Through this process, I identified 62 reports that did not meet criteria for inclusion in the sample, including reports that were not publicly available (23) and references which were not reports such as books and PowerPoint presentations (39). I also identified one additional duplicate report. After removing these references from the dataset, the final analytic sample included 215 reports, with 46 (21%) referenced once and 169 (79%) referenced more than once. The missing or reclassified reports identified during the search process reduced the sample equally for reports referenced once and those referenced more than once, as reported in Table 3.1.

Table 3.1

Reports Included in the Analytic Sample

Frequency of use	Initial sample	Removed	Analytic sample
Total	278	63	215
Once	215 (79%)	50	169 (79%)
More than once	59 (21%)	13	46 (21%)

Note: Removed reports were not publicly available, were not reports including books and PowerPoint presentations, or were duplicates.

Data Visualization Checklist

I used the DVC to rate the quality of the data visualizations in reports. The DVC is a unidimensional rating scale developed by Evergreen and Emery in 2014 and revised in 2016. A copy of the DVC is included in the Appendix. The scale consists of 24 performance statements addressing the five aspects of high-quality visualizations that I described in Chapter 2, including text, arrangement, color, lines, and overall presentation. For example, one performance statement is "Data are labeled directly." Users of the tool score each statement as fully met (2), partially met (1), not met (0), or not applicable (removed from total available points), for each data visualization in a report. For example, a rater would score the guideline "Axes do not have unnecessary tick marks or axis lines" as not applicable for a pie graph. A report with no data visualizations will obtain a score of 0 on the checklist. Otherwise, each report was awarded a final DVC score based on the average of the DVC scores awarded for each data visualization in the report, discussed further in the Analysis section.

Raters

The data visualizations in the identified reports were rated using the DVC by me and volunteer raters. I recruited raters in-person at a data visualization session at the American Evaluation Association's (AEA) annual conference. In addition, Evergreen presented the study as an option to participants in her data visualization workshops held in spring 2017 in Honolulu, Hawai‘i. For both, we promoted the study as an option for individuals interested in data visualization to become more acquainted with the DVC. I did not employ a selection criteria when identifying report raters. Anyone who expressed interest and completed the mandatory rater training was invited to serve as a rater. Not all potential raters were evaluators, although all

were involved with creating or using data visualizations in their work.[6] A total of 39 individuals initially indicated an interest in the study. Of these, 19 confirmed participation and 16 completed the rater training. One rater was a colleague who had not attended the AEA session or data visualization workshops but who was interested in learning more about the DVC. He only partially completed the rating assignment. Due to the partial completion of the rating, his scores were removed from the final analysis. Including myself, a total of 16 raters assessed the data visualizations in the reports.[7]

Rater Training

All raters completed a one-hour training webinar (live or recorded) before receiving their assigned reports. The training consisted of a slide show with 59 slides of images and text providing

- what is and is not a data visualization;
- what to include in the DVC ratings (ratings are based on graph titles, notes, and content and do not include narrative from the report);
- a general orientation to the checklist;
- examples of common mistakes to avoid when rating including giving lower marks when the performance statement was met because raters felt it could have been better, and giving higher marks when the performance statement was not met because raters gave the visualizations the benefit of the doubt;

[6] One of the changes Evergreen and Emery made in their 2016 revision of the tool was removal of language that described the Data Visualization Checklist (DVC) as a tool exclusively for evaluators (Evergreen, personal communication, January 3, 2017)

[7] To ensure I was not biased in my ratings, I assigned each report an ID and kept the rating data separate from metadata about the report, including the frequency of use variable. As such, I did not know which reports were used more than others until after the rating process was complete.

- examples of when to rate a statement “not applicable”;
- criteria for “met”, “partially met”, and “not met” for each of the 24 guidelines; and
- example graphs to illustrate application of each guideline.

I also included additional call out warnings and/or tips in the training slides for nine of the 24 guidelines based on common errors or points of confusion identified during cognitive interviews. For example, I included the text, “Warning: This is one people commonly rate higher because they don’t mind tick marks” with the guideline “Axes do not have unnecessary axis lines or tick marks.” Each rater was also provided a hard-copy of the training slides along with their assigned reports to use as reference during the rating process. The training was essential, as it provided raters guidelines and examples to follow when applying the DVC to ensure common interpretation of the performance statements.

Rating

Individuals who attended a live webinar received a rating packet including (a) their assigned reports, (b) a rating template, and (c) a PDF copy of the rater training slides via email following the training. All others received their rating packet after I received an email confirmation that they had watched the training video. Each rater received no more than ten reports to rate and each rater independently rated their assigned reports. Raters were given three weeks to complete the ratings and all ratings were completed within six weeks. Upon completion, raters were given a Starbucks gift card worth $5 as a mahalo.

Alternative Explanations for Use

The theory driving this study is that data visualizations in reports aid readers' ability to make sense of information presented, thus increasing their likelihood to use the information. However as mentioned earlier, use is a sticky spider web, and there are many potential factors contributing to the likelihood that an evaluation will be used (Alkin & King, 2017; Cousins & Leithwood, 1986; Johnson et al., 2009; Leviton & Hughes, 1981). Drawing from prior research, I explored alternative explanations for use including, (a) if the report was more similar to traditional or advocacy research as a proxy for credibility, (b) if the user was academically trained or not, and (c) the length of a report. I describe each in more detail below.

Type of Report

The type of report refers to if a report follows the norms and content of a traditional research report or not. In Evergreen's (2016) work-in-progress, she found that reports which were highly designed, for example magazine quality, were less used, although the finding was not significant. She suggested that one possible explanation for the finding was that some readers found the well-produced reports less trustworthy or rigorous than those which followed norms for academic research. Prior research on evaluation use also found credibility to be a factor in use (Cousins & Leithwood, 1986; Leviton & Hughes, 1981).

In a similar vein, Reckhow et al. (2015) categorized reports referenced in legislative testimony as *traditional research* or *advocacy research* based on eight criteria.[8] These included; who produced it, discussion of caveats, inclusion of policy recommendations, the production

[8] Reckhow et al. (2015) used the term *policy research* but referred to this in their text as *traditional research*. The authors work and publish in the field of policy research, where the term policy research refers to traditional academic research. Therefore, I use the term *traditional* research throughout the manuscript for clarity and to reduce confusion.

quality of the report, the presentation of evidence or data, inclusion of references, tone of objectivity, and explanation of study methods. Ultimately, the concept of categorizing reports as advocacy or traditional research gets at the concept of legitimacy, where advocacy research is considered less legitimate or trustworthy.

Using Reckhow et al.'s (2015) coding schema, I scored reports on eight criteria—producer, conclusions, policy recommendations, production quality, evidence, citations, tone, and methods—using a three-point scale. For each criterion, reports received zero points if they were more like traditional research and two points if they were more like research published by advocacy think tanks. Work that contained aspects of both traditional and advocacy research for a specific criteria received one point. For example, a work co-authored by a university professor and by a partnering organization staff member would receive one point for the "producer" category. After scoring reports based on each of the eight criteria, I totaled the points to generate a report type score that ranged from 0 to 16 where reports with a report type score closer to zero were more like traditional research, and reports with a report type score closer to 16 where more like advocacy research.

Coding. Working with a second coder, I used Reckhow et al. (2015) coding scheme to classify the reports in my study as either more or less like traditional or advocacy research. We went through three rounds of coding five reports to reach 79% agreement in our ratings, at which point we proceeded to code the remainder of the reports in the sample. By a stroke of luck, five reports in my analytic sample had been included in Reckhow et al.'s study, and the authors were generous enough to share their original coding with me. Therefore, in the first round, we coded the five shared reports and reviewed our agreement between each other and the original ratings. When there was disagreement in our ratings, we referred to the original coding awarded by

Reckhow et al. for resolution. Based on this process we added additional descriptive language and examples to clarify Reckhow et al.'s original coding scheme (see Appendix).[9]

Academic Affiliation of the User

One of the problem statements driving my study is that individuals not familiar with academic writing and research may be less able to, or less interested in, making sense of information in research or evaluation reports which follow academic publication norms. For example, research on the interpretation of graphs found that readers with more expertise in a topic had an easier time with and were more accurate in their interpretation of information presented in graphs (Tsuji & Lindgaard, 2014).

Along with expertise, where someone works may inform their receptivity to information presented in a report (Newman, Brown, & Braskamp, 1980). In particular, some contexts may place greater value on following academic publication norms than others. For example, someone in a university setting may feel more pressure to cite works published in refereed journals where someone in a non-university setting may not. For these reasons, I included academic affiliation of users as an alternative predictor of use where academic affiliation was based on a user's affiliation with an institution of higher education at the time the reports were used. The original data from Reckhow et al. (2015) included the organization for each witness who submitted testimony which they grouped into eighteen categories, for example "government", "non-profit", etc., including "university." I used these codes to classify the academic affiliation of the user as either university-affiliated or not university-affiliated.

[9] The original coding scheme published by Reckhow et al. (2015) only included two codes; traditional or research. However, the authors rated the reports on a three-point scale: 0, 1, and 2. Following this, we used the five example reports in the first round of coding to develop descriptive text and examples to further distinguish when a report would receive a 0, 1, or 2 for each category.

Report Length

Everyone is short of time, and this is particularly true for policymakers tasked with making sense of multiple sources of information in order to make decisions. It is a fact that lengthy reports take more time to read. Thus, I am interested if the length of a report plays a role in the symbolic use reports based on the assumption that policymakers may be more likely to read shorter reports or report summaries, and thus, more likely to use the information included in the report.

Analysis

I conducted two regression analyses to answer the questions if the use of data visualizations and the quality of data visualizations predicted the likelihood reports would be used more than once. I used frequency of use as the dependent variable in both regression equations, discussed further below. In addition, I used the same alternative predictors of use in each equation; type of report, user affiliation, and length of report (in pages). More details about the dependent and independent, or predictor variables, as well as the regression analyses follow.

Dependent Variable

The frequency of use was a count of how many times a report was referenced in congressional testimony related to teacher quality between the years 2000 to 2015. Originally I transformed this into a dichotomous variable where *0* = the report was used once and *1* = the report was used more than once. However, after determining the distribution of the variable matched a Poisson distribution, I recoded the data so that a report referenced once = *0*, a report referenced twice = *1*, reports referenced three times = *2*, and so forth.

Independent Variables

My primary independent variables were the percent of data visualizations in each report and the DVC percent score for each report. Each are described below, along with the alternative predictors included in the model.

Percent of data visualizations. I calculated the percent of data visualizations based on the total number of data visualizations in a report divided by the total number of pages in the report including any appendices.

DVC percent score. I calculated a DVC percent score for each graph based on the total points awarded divided by the total points possible, minus items that were scored as not applicable.

Total points/ (Total points possible – Total number of points possible from items scored "N/A")

An example of an item scored as not applicable would be rating axis tick marks on pie graphs. I calculated the DVC percent score for the five graphs included in the interrater reliability analysis based on the average across the 14 raters (see the *Interrater Reliability of the DVC* section). For all other graphs, the DVC score for each graph was generated by a single rater. Because the unit of analysis for the regression analyses was the report, I then generated an overall DVC percent score for each report based on the average DVC percent score for all graphs in that report.

Type of report. The type of report score was based on total points awarded across eight categories related to classifying a report as more like traditional research or more like advocacy research. Reports could have a total score between zero, indicating the report mirrored traditional research, and 16, indicating the report mirrored advocacy research.

User affiliation. Some reports were referenced by both individuals affiliated, and not affiliated, with a university. Initially I calculated the average user affiliation for reports which had more than one witness. For example, a report referenced by one person affiliated with a university and one person not affiliated, the user affiliation score was 0.5. However, there were relatively few cases of reports with mixed user affiliation. Due to concerns with small cell sizes, I transformed the data into a dichotomous variable. Reports referenced by at least one witness affiliated with a university were classified as "university", coded as 1, and all other reports were classified as "non-university", coded as 0.

Length. The length of report was the total number of pages in the report, including appendices.

Regression Analyses

I elected to conduct a Poisson regression, which is appropriate when the independent variables may not be normally distributed and the outcome, or dependent variable, is a rare occurrence (Azen & Walker, 2011). To explore if the *use* of data visualizations predicted the frequency that the reports were used, I conducted a Poisson regression with frequency of use as the dependent variable and the percent of data visualizations (viz) as the primary predictor with type of report (type), user affiliation (aff), and report length (length) as alternative predictors. The regression equation for the first analysis follows.

$$\gamma = \beta_0 + \beta_{viz} + \beta_{length} + \beta_{type} + \beta_{aff} + \varepsilon$$

To explore if the *quality* of data visualizations predicted the frequency that the reports were used, I conducted a second Poisson regression with frequency of use as the dependent variable and the

DVC percent score (dvc) as the primary predictor with type of report (type), user affiliation (aff), and report length (length) as alternative predictors. The regression equation for the second analysis follows.

$$\gamma = \beta_0 + \beta_{dvc} + \beta_{length} + \beta_{type} + \beta_{aff} + \varepsilon$$

Construct Validity of the Data Visualization Checklist

The present study is the first time the DVC was used for empirical research. As such, prior to using it for the rating study I conducted cognitive interviews as well as an interrater reliability (IRR) analysis to establish evidence the tool was a reliable measure of data visualization quality. In this section I present a brief overview of the tool's development and my concerns using it, before describing my approach to the cognitive interviews and the resulting evidence of construct validity. I conclude the section with a description and results of the IRR analysis.

Development of the Data Visualization Checklist

The DVC was developed in 2014 based on research and on the authors' experiences training clients to improve data visualizations (Evergreen, personal communication, January 3, 2017). At the time it was developed, the DVC was vetted with six practicing evaluators actively involved in evaluation reporting (Evergreen, personal communication, January 3, 2017). The checklist was revised in 2016 for clarity including updating the descriptions of four elements. No performance statements were changed (Evergreen, personal communication, January 3, 2017).

In an exhaustive history of content validity, Sireci (1998) summarized the consensus that validity must be considered in terms of the purpose of the measurement. The original purpose of the DVC was for evaluators to self-assess and improve the quality of their data visualizations in

reports (Evergreen, personal communication, January 3, 2017). As such, review of the tool by evaluators engaged in evaluation reporting provided evidence of content validity for the original intent of the checklist (Gulliksen, 1950).

Validity Concerns

Unidimensional rating scales are used to measure a single underlying concept and are good when used to measure something where you can have more or less of it, for example, depression (Bagby, Ryder, Schuller, & Marshall, 2004). However, the DVC is intended to measure the presence of distinct elements necessary for quality visualizations, which may be more nuanced than a single concept scale. Also, the language in the DVC can be ambiguous, for example, "The graph has an appropriate level of precision." In addition, using a rating scale very similar to the DVC, Evergreen (2011a) found just 3 of 23 items, about 13% of the scale items, had an acceptable inter-rater reliability of .60 or above (Landis & Koch, 1977). For these reasons, I collected additional evidence of construct validity through cognitive interviews.

Cognitive Interviews

One way to collect evidence of validity is to determine to what extent response processes follow the intent of the measure, or in other words, how well raters' interpretation of a measure aligns with its intended use (AERA/APA/NCME, 1999). Cognitive interviews (CI) can be used to explore response processes, as they allow researchers to reveal the cognitive processes or steps people think through when responding to a tool or measure (Willis, 2005). CIs consist of structured and probing questions that prompt individuals to speak aloud their internal processing while answering a survey or completing a measurement tool. The CI was born out of survey research and continues to be predominately used in survey development (Sjetne, Iversen, &

Kjøllesdal, 2015; Willis, 2005), but it can be used for any instrument that requires more than a simple stimulus/response sequence and has been used with unidimensional scales (Tomlinson et al., 2016).

***Interview structure*.** CIs are often conducted in rounds as an iterative process for the purpose of revising and re-testing a tool (Willis, 2005). However, the primary purpose of CIs in the present study was to collect evidence of construct validity. As such, I conducted one round of interviews. The CIs lasted approximately 45 minutes, were audio-recorded, and took place at a time and location that was convenient for participants in a space with minimal background noise. For example, many interviews took place in office conference rooms.

Interview protocol. The two most common techniques used within cognitive interviewing are think-alouds and probing questions. Willis (2005) noted that both techniques tend to bleed into each other in practice. For example, interviewers probe participants during think-aloud exercises and vice versa, participants tend to talk through their thinking during a probing interview. Based on the opportunities and limitations of both approaches, I used a modified think aloud protocol, described below, with the addition of selected probing questions for the following reasons:

- The think aloud protocol is open-ended and can accommodate unanticipated challenges in use of the DVC not identified a priori.
- Probing questions allow the interviewer to focus on particular elements or areas of concern within a tool (Willis, 2005).

I used a modified think-aloud approach with the use of structured probing questions for one section in the DVC that consisted of global statements anticipated to be problematic for

users. In traditional think-aloud exercises, a researcher directs participants to think aloud while attempting to solve a problem (Ericsson & Simon, 1993) which is often a question or item on a survey or other tool. In the modified approach, I provided the participant a sample data visualization (see Appendix), read a performance statement from the DVC aloud (e.g. "Data are intentionally ordered") and the participant walked through their thought process to get to a rating score.

***Probing questions*.** Throughout the think-aloud, I used reactive probing questions to follow-up or dig deeper as needed based on the responses and behaviors of the participants. For example, if a participant had difficulty rating an item, I followed-up with the question, "What was going through your head as you tried to rate that item?"

In the last section of the DVC, there are four global statements intended to provide holistic ratings of each data visualizations—for example, "The graph highlights a significant finding or conclusion." For these four items, I asked each participant the same set of standardized questions in order to document potential problems they may have with the technical language or the complexity of the statements (see Appendix).

***Interviewer and participant training*.** Interviewers trained in field interviewing are able to conduct CIs without the need for additional training (Willis, 2005). However, CI participants typically do require some training in order to participate in the think aloud section of the interview. The training that I provided was embedded into the interview and consisted of (a) a standardized script read to participants at the beginning of each interview describing the think aloud exercise and expectations, and (b) a practice exercise conducted prior to beginning the think aloud prompts. The introductory script and practice exercise are included in the Appendix.

The second purpose of the CIs was to identify potential problem areas, for example challenges with ambiguous or technical terms that could impact reliable application of the tool in order to address these issues during the rater training.

Participants

The recommended number of interviews for any one tool is between 5 and 15 (Willis, 2005). In line with this recommendation, I recruited nine participants in person and via email from two data visualization workshops held on the island of ‘Oahu in spring and summer 2017. Prior studies that have used CIs to establish evidence of validity have fallen in this range (Bailey, Tully, & Cooke, 2015). Ideally, participants in CIs are representative of the population intended to use the tool under consideration (Willis, 2005). As such, I recruited CI participants from data visualization workshops with the understanding that they would be interested in data visualization, and thus, constitute the population likely to use the checklist.

As part of the interview process, I collected additional background information from the participants, including prior exposure to the DVC, prior use of the DVC, self-assessed familiarity with the DVC, prior experience creating data visualizations, and length of experience in years. The survey is included in the Appendix, and Table 3.2 provides a brief summary of key participant characteristics.

Although there was some variation in participant characteristics, more participants had seen the DVC before but had not yet used it, and most rated their familiarity with the checklist in the middle. Also, most participants had previously made data visualizations and had more than 5 years of experience making them. The background characteristics of the interview participants indicated I did not get the perspective of individuals who had never heard of the DVC before, nor

those who felt comfortable enough with it that they could teach others. Also, the perspectives represented by the interviews are skewed to those with more experience with data visualizations.

Table 3.2

Characteristics of Cognitive Interview Participants

Characteristic	Count
Have you seen the DVC before?	
Yes	5
No	4
Have you used it before?	
No	6
Yes	3
On a scale from 1-5 how familiar are you with data visualization, where 1 is not at all and 5 is you could teach it?	
1	0
2	2
3	5
4	2
5	0
Have you previously created data visualizations?	
Yes	7
No	2
If yes, how many years have you been creating data visualizations?	
5+ years	5
1 to 4 years	1
Less than 1 year	1

Analysis and Results of the Cognitive interviews

Data from the CIs consisted of interviewer notes collected during the interview, reflective field notes completed immediately following an interview, and audio-recordings of each interview. Following each interview, I listened to the audio-recording and further developed the interviewer notes and reflective field notes into a rich description of each interview. The rich

description—based on a composite of interview notes, field notes, and a summary of the audio-recording—constituted the data used for analysis.

I used the constant comparative method, following an adapted grounded theory approach, to identify common themes across interviews. While a grounded theory approach to analysis is conducted without a priori assumptions or theories about the structure or nature of the data (Strauss & Corbin, 1990), an adapted grounded theory utilizes known information to assist in initial generation of categories for data analysis. This approach has been used by researchers to establish evidence of validity (Brod, Tesler, & Christensen, 2009). For the DVC study, I based initial codes on foundational themes within the DVC in order to explore supporting and contrasting evidence in the data related to those themes. Before analyzing the data and developing the rater training, I confirmed my understanding and interpretation of the DVC items with Stephanie Evergreen, co-creator of the tool, via a phone meeting on June 30, 2017.

Key concepts within the DVC. The DVC consists of 24 statements grouped into five categories; text, arrangement, color, lines, and overall design. The tool also includes summary statements for each category to provide users the gist of what the category aims to achieve. These summary statements are listed below.

- Text should support the takeaway message and be formatted to grab readers' attention.
- Thoughtful arrangement of graph elements (proportions, axes, order, etc.) aid in readers' interpretation of the data.
- Color use should be accessible and deliberate.
- Lines create noise and should be muted or removed.
- Data visualizations should be used to deliver a takeaway message in the data.

Drawing from these summary statements, as well as the research on cognition, evaluation communication, and information science, I reviewed the guidelines and determined that each guideline aimed to either improve readability or aid in interpretation of the graph. Following this, I analyzed the interview notes and coded participants' responses for two things; alignment with the underlying concepts of readability and interpretability, and the level of difficulty participants had rating each guideline. An overview of each guideline, the key concept the guideline supported, representative comments from participants demonstrating alignment—or lack thereof—to the concept, and the difficulty rating for the guideline are all presented in Table 3.3. For guidelines that presented some difficulty for interviewees, I also included a summary of how I addressed the difficulty in the rating training. This is presented in italics under the guideline statement.

Difficulty rating. Drawing from (Tomlinson et al., 2016) I rated the difficulty participants had applying each guideline on a 3-point scale where *0* = no difficulty, *1*= the participant requested helper language and/or struggled to make a decision, and *2* = the participant remained confused after seeking clarification. Difficulty ratings were based on participants' challenges with the guideline and not on difficulty interpreting the graph. Also, I did not consider the time it took participants to rate the graph. Across the 24 guidelines and nine interviews, there were only two instances an interviewee remained confused after seeking clarification, therefore I report difficulty level as either "low" if one or fewer participants had difficulty, or "some difficulty" if more than one interviewee sought clarification or struggled to make a decision.

Alignment to key concepts. Across all guidelines, with the exception of "*Graph has an appropriate level of precision*" discussed further below, participants' responses demonstrated alignment to the underlying concepts of either improving readability or aiding interpretation.

Often, responses included literal reference to the language in the guideline. For example, most participants awarded partial points for the guideline *"6-12 word descriptive title is left-justified in upper left corner"* and in getting to that rating they specifically called out the number of words in the title and the title's position on the page (See example comments in Table 3.3).

The participants also demonstrated alignment with the broader concepts of readability and interpretability. For example, they explicitly referenced the readability and/or interpretability of the graph when applying the ratings (e.g. "Well, you have to spend a lot of time looking at the graph to know what the graph is saying, it's not in the title."). Also, these instances largely aligned with the key concept for each guideline (See example comments in Table 3.3).

Non-alignment or Misalignment to Key Concepts. Across the nine interviews, the participants described how they rated 24 guidelines, resulting in a total of 216 opportunities to demonstrate alignment to and understanding of the tool. Of the 216 applications of the tool, there were 37 instances (17%) where participant statements did not align with the guideline and/or the key underlying concept of the guideline. Of note, eleven of these came from Rater 6 who individually provided 5% of the overall instances of misalignment. This is discussed further in *The Outlier* section below.

After coding the full dataset, I reviewed the instances of misalignment to look for patterns and explored if the disconnect was due to misunderstanding of key concepts, or due to challenges with the tool—for example use of ambiguous or overly technical terms. All instances of misalignment, save the unique case of Rater 6 discussed below, were due to ambiguity in the checklist and were addressed by clarifying terms and providing examples in the rater training. How each of these were resolved in the rater training is included in italics under the guidelines column in Table 3.3.

The guideline, "The graph has an appropriate level of precision," was the most challenging for participants. Five of the nine participants' demonstrated difficulty rating this guideline and all five misinterpreted the statement due to ambiguity in the concept of "precision." In particular, the participants were not sure what aspects of the graph to look at in order to rate its precision. One participant summarized it well, "*I guess what I was thinking about… we're talking about precision like, what specific level of precision are we talking about? Am I looking at the graph as a whole or am I just looking at the data points?"*

Anticipating difficulty with this question, I had asked all of the participants to tell me in their own words what they felt the guideline meant and most discussed it as providing the right amount of information. For example "*That it's, it's kind of a goldilocks – that it's just enough data that lets me understand the story or patterns being shown but not so much that it convolutes that."* A few also discussed the concept in terms of simplicity or clarity: "*Expressing with the closest degree of simplicity what's represented by the data*." These concepts align with the key concept of readability associated with this guideline, which suggested that providing examples and text to clarify which aspects of the graph to rate may resolve the issue.

Even in error, there was alignment. When the participants misunderstood or incorrectly applied the guidelines, errors were related to the overall concepts of readability and ease of interpretation. In the two examples below, Rater 3 gave the sample graph a score of "1" for the guideline "Text size is hierarchical and readable" and Rater 8 gave the sample graph a "1" for the guideline a "6-12 word descriptive title is left-justified in upper left corner."

"*It's an ineffective heading because it's outlined. So I'm rating it based on the effectiveness of the headline for a chart. Visually without even going into whether it's accurate – it's just challenging to read… [Score?] One, partially met*."

"*I think it's hierarchical, the subtext is smaller. Um… although I'm not sure about the columns labels underneath, they look smaller, so I would say it's hierarchical. In terms of readability, I would say it's not very readable but yeah I forget what that is called but like outline letters so I would say one, partially met."*

In both instances raters referred to something other than the element identified in the guideline to make their rating and therefore, did not align to the guideline. However, their thought processes aligned to the underlying concept of readability. What these example cases demonstrate is that participants adhered to the key underlying concepts, even when individual guidelines were incorrectly applied. While this is helpful for determining evidence of construct validity, the examples highlight potential challenges in inter-rater reliability if the sources of error were not addressed.

The outlier. One rater consistently interpreted the rating guidelines differently than the other raters and often misinterpreted the intention of the guidelines. For example, when asked to rate the presence of unnecessary tick marks or axis lines, the rater responded, "*Axis lines? Not applicable. It's not a line chart, it's a bar chart, and I don't see any access points. [Axis points?] You think of dots and straight lines*." I considered that the misinterpretation was due to the rater's lack of experience and/or exposure to data visualizations. However, the rater had similar background characteristics, familiarity with the checklist, and experience as other raters who correctly interpreted the checklist items. This suggests that the rater's misinterpretation of the guidelines was idiosyncratic and representative of the individual variation we are likely to see in the population intended to use the checklist. We see further support for this in the inter-rater reliability estimate for a single rater which was lower than the estimate for average measures across a group of raters, discussed below.

Interrater reliability of the DVC

A key aspect of validity is reliability – do raters apply the guidelines in similar ways that allow us to trust the results? I looked at the interrater reliability (IRR) of data collected with the DVC using intra-class correlation (ICC; Shrout & Fleiss, 1979). ICC is an appropriate measure of IRR for interval or ratio data with two or more raters (Hallgren, 2012). I selected a random sample of two reports containing five distinct data visualizations (n = 5) to assess reliability of the DVC using a fully crossed design; that is, each rater (k = 14) rated each data visualization (Hallgren, 2012). Of note, two of the 16 raters joined the study late and their scores were not included in the reliability analysis. I selected a two-way consistency average measures ICC with mixed effects because the same raters rated the same reports (two-way), raters were not randomly selected but reports were (mixed), and I was interested in the consistency of ratings across raters rather than absolute agreement (consistency). Also, I wanted to know the reliability across a group of raters rather than the reliability of a single rater (average measures).

I ran the ICC analysis using SPSS version 22 and found the ICC (2, 14) = 0.87, CI = (0.73, 0.95), based on a 95% confidence interval. ICC values between 0.75 and 0.90 are considered to have good reliability (Koo & Li, 2016). Of note, the ICC estimate for single measures— individual raters—was lower than the average across a group of raters (0.58). ICC values between 0.50 and 0.75 are considered to have moderate reliability (Koo & Li, 2016). However, for the purposes of my study, the results of the IRR analysis indicates the tool has good reliability.

Reliability estimates are based on ratings administered after the raters had completed the required training, which addressed ambiguity in the guidelines identified during the interviews.

For this reason, the IRR estimates reported here are applicable only for raters who have completed a similar training.

Table 3.3

Alignment of Participant Comments to DVC Concepts and Difficulty Level, by Guideline

Category	Guideline	Concept	Example Comments	Difficulty
Text	6-12 word descriptive title is left-justified in upper left corner	Interpretation	"I would say one, partially met because it looks like it's more than 12 but does meet the left justification."[a] "Well, it's left justified at the top. Well, I guess I would give it .5 because it's so many words. It's not descriptive really. Well, it's describing the content but it's not making the point that's made obvious in the visualization. Well you have to spend a lot of time looking at the graph to know what the graph is saying, it's not in the title."	Low
	Subtitle and/or annotations provide additional information	Interpretation	"There is a subtitle that provides additional information and I'm not sure how helpful the additional information is but since that wasn't what the statement asked I guess I would say it met it." "Okay so there is a subtitle, it's.... I mean I guess there's additional information…I'd say that's fully met though because it is, without having to look at the visualization you know that the graph will tell you race and ethnicity and the years that you may or may not be interested in."	Low
	Text size is hierarchical and readable	Readability	"Umm… Yeah I think because of the fact that I've created things like this before, I can tell that it's hierarchical but that it's not, at a quick glance to the naked eye it's really close – the difference between 16 and 14 so I can really see this is the most important thing and the rest follow suit as it goes down it looks the same."	Low

Category	Guideline	Concept	Example Comments	Difficulty
			"Yes, I would say I'd give that a two although it's hard to tell and the only thing I might change from a layout perspective it looks like a white letter with a black outline so it's readable but visually distracting to have a non-standard type. A two."	
	Text size is horizontal and readable	Readability	"I would say all the text looks to be horizontal so two, fully met. Yeah, I don't see any vertical or diagonal text." "Yes, text is horizontal. I just looked at it and it is horizontal. I am looking at the title, the additional information on the data, the dates, the legend, the percent numbers and the categories on the x-axis."	Low
	Data are labeled directly *The helper language for "Data are labeled directly" calls out the use of legends as an example of not directly labeling data. All interviewees missed this distinction. This was addressed in the rater training.*	Interpretation	"So I am looking at the graph itself at the percent and categories along the x-axis and I would say yes, it's directly labeled because it's labeled not only on the axis but because each bar is labeled." "Yes, the data are labeled directly they're over each piece of the bar graph there's a label and they've tacked on little subcategories to each thing and there's a legend so I guess yes, fully met"	Low
	Labels are used sparingly *Some raters did not mind redundant labels. This was addressed in the training by flagging raters to be wary of giving this guideline a higher rating because they didn't mind the redundancy.*	Readability	"No, everything is labeled. Labels are not used sparingly… so I don't know, I hate giving 0s to things, but I can't image any way you could have labeled it more so I have to give it a 0." "[Laugh] That seems a little contradictory to the previous, and 'sparingly' is sort of subjective. To me I think it's just the right amount of labeling. What's the point of having that statement if you want it labeled? To make it understandable and not	Low

Category	Guideline	Concept	Example Comments	Difficulty
			crowding? If that's what they mean by that statement, because it [the labels] provides enough info but doesn't overwhelm."	
Arrangement	Proportions are accurate *Raters struggled with what to look at and how to judge accuracy. This was addressed in the training by pointing out what to look at and providing examples.*	Interpretation	"I'm looking at what they mean by proportions, I guess the bars. Since it lines up with gridlines I would say that….and the bars are all the same size, looks like I would say yes, fully met. [Unsure?] Because I still don't know proportions – Are they talking about the bars or are they talking about all the things on the page? Like all the different elements of the key and the title but, Yeah but if just the bars, yes I would say fully met." "Accurate to what? So if we're talking about a 34% bar is taller than a 32% bar and I'm looking at each one and then along the y-axis so I would say fully met, comparing the spaces and where the bars land and how they're labeled.	Some difficulty
	Data are intentionally ordered *Raters struggled to assess intentions. This was addressed in the training with examples of intentional ordering.*	Interpretation	"By year makes sense and that's appropriate because it's chronologically, but I can't figure out why white, black and Hispanic are in the order they are. The white is the lowest and then it gets progressively larger but that seems like an odd way to do it that way" "If I'm looking at this and using the legend that shows by year and years are kept in the same order, I guess? It's hard to say because it's hard to see the difference between 1983 and 1987 they're both coming out to white, but I guess they are in the same order for each sub-group at the bottom so I would say 2, fully met because there is an order and it's chronological.	Some difficulty
	Axis intervals are equidistant *Difficulty came from people not knowing what an axis*	Interpretation	"Yes – I would give that a 2. The spacing looks even to me. [What are you looking at?] Everything really, the spacing between the bars, the categories, and the spacing between the y-axis lines."	Some difficulty

Category	Guideline	Concept	Example Comments	Difficulty
	interval was. This was resolved in the training with images.		"Okay my immediate thought is what do you mean by Axis intervals? – Certainly the ah… along the axis they are equidistant, and I'm assuming that's what you mean so I would say yes, fully met, if that's what that means. Fully met."	
	Graph is two-dimensional	Readability	"So yes, the graph is 2D. I'd say fully met, 2. [Tell me a little more about how you got to your rating.] Well the graph is 2-dimensional because there's no 3-D bars sticking out, it's just flat. Even the words, they're still 2-D, there's no background shading, everything is 2D."	Low
	Display is free from decoration	Readability	"Okay well, there's no pictures or like flowers or icons or anything but I mean some people might say that the title is decorative so, I'd say partially met because the title looks decorative to me." "Okay so again immediately I'm trying to think okay "decoration", um I'm assuming decoration means that it's ornamental in nature and it is free of that, so I would say fully met there's nothing extraneous on here and nothing decorative."	Low
Color	Color scheme is intentional *Raters struggled to assess intentions. This was addressed in the training with examples of intentional coloring.*	Readability	"Ah, Yes it appears to be intentional though I don't know what that is. I'm talking about the black lines in the 1972 bars where obviously it's calling out 1972 for whatever reason, trying to highlight that." "The color scheme is intentional… I would say partially only because I don't see a difference between 1983 and 1987 because they're both white and if you're making a distinction than one of those should have a different color."	Some difficulty [b]

Category	Guideline	Concept	Example Comments	Difficulty
	Color is used to highlight key patterns *Raters struggled to assess the use of color to highlight patterns. This was addressed in the training with examples.*	Interpretation	"Patterns? … ah, I'm not sure I see the pattern so I'm not sure it's applicable. I see it might be highlighting 1972 but I'm not sure what the pattern would be in it. "In this case because 1972 is the only one with color I would say it's been used to highlight that year, and so maybe it was intentional that they wanted us to look at 1972 and not focus on the other two years….Yeah and in each case, 1972's the highest so. I don't really know what that's trying to show us though."	Some difficulty
	Color is legible when printed in black and white	Readability	"Well, I mean maybe partially I have to say that again because I can tell the difference between the dark, between the 1972 and the 1983 and 87, but between 1983 and 1987 I really can't tell a difference so partially I can tell a difference when it's printed in black and white so I guess I have to give it a one." "Um… so there is a color there that got... So it does not, it is not legible. I can read the words but the data cannot be read correctly because it doesn't print correctly in black and white, presuming it was originally in color."	Low
	Color is legible for people with colorblindness	Readability	"So, I think people with color blindness can read these contrasting colors white and black so yes, fully met, 2."	Low
	Text sufficiently contrasts background	Readability	"I'd give that a one going back to the title the white font with the black stroke doesn't contrast too well on the white background. It doesn't stand out as clearly as it could, as the text gets smaller it's harder to read." "I can read all of the text so I think that's an accurate statement but again, outlined fonts are challenging to read period, it's not pleasant to read. One."	Low

Category	Guideline	Concept	Example Comments	Difficulty
Lines	Gridlines, if present, are muted	Readability	"Not met, 0 because there are gridlines and they stand out pretty strongly they are the same width as the lines around the bars… and they seem to have this weird pattern too, it's not even just one thin line, it's like thick and thin and kind of extra distracting."	Low
	Graph does not have border line	Readability	"So the graph does have a border line, yeah, all around the entire graph and both axes so, 0 not met."	Low
	Axes do not have unnecessary tick marks or axis lines *Flagged raters to be wary of giving this guideline a higher rating because they didn't mind tick marks or axis lines in the graph.*	Readability	"No, they have unnecessary tick marks on the y-axis and I get on the one hand they're like, let's be exact, let's put them in so people can see but we don't need those , I would say 0." "Um.. okay so I'm looking at the axes and I don't think there's unnecessary markings there so 2, fully met. I in fact like the intermediary lines between the major point because it makes it easier to see the bars going up and I think the tick lines along the x-axis but I think they help delineate the different sets of % I think it helps make it more clearer."	Low
	Graph has one horizontal and one vertical axis	Interpretation	"Yes, yes it does. Fully met 2. You've got the y-axis, you've got the x-axis and if there's more then I'm not seeing them. Maybe I'm not understanding the statement so, that one seems fairly straightforward."	Low
Overall	Graph highlights significant finding or conclusion *Called out helper text for raters to consider if the data are worthy of graphing rather than assess if the graph provided a takeaway message.*	Interpretation	"I feel like there was intentional highlighting and with a little bit of thought you could get there but the wording of the graph didn't tell me what to look for and it took me some time for me to get to what I think the conclusion should be." "My initial thought was this is just descriptive data so there's a lot you can pull out of it. Then I thought why don't I look at it and analyze it. It didn't like jump out at me. "	Some difficulty

Category	Guideline	Concept	Example Comments	Difficulty
	The type of graph is appropriate for data *Flagged raters to be wary of giving this guideline a lower rating because they think it could be done better and pointed to resources which provide an overview of appropriate graphs for specific types of data.*	Interpretation	"I think if you're using this data you need to figure out why you're using it and if you're trying to highlight a specific point there would be a different kind of graph you could use – I'd have to think which one, but I think there's other ways to go with this. Like you could use, I don't know trend lines, or some kind of horizontal graph, I don't know." "So I'm thinking it's not inappropriate. Like a bar graph or a column graph, it's showing a percent of a whole, so I think this works."	Some difficulty
	Graph has appropriate level of precision. *Raters struggled with which graph elements to assess when rating precision. This was addressed in the training by pointing out what to look at in the graph and providing examples.*	Readability	"I don't really know what that means. I mean it's clear there's data points on here, there's a definite time period but the time intervals aren't even. I would give that a one, partially met because the exact percentages on here but the uneven time intervals." "I guess so… I mean there was an attempt made at precision because they labeled every single data point and that's good, I guess, from a precision point. My only hesitation is I don't know if there was another way to be more precise about how these numbers are presented rather than presenting them in a legend."	Some difficulty
	Individual chart elements work together to reinforce the overarching takeaway message	Interpretation	"Yeah – I'd just say one partially met. Because they're working together to give me the message, but then the message isn't really explicit because the title could play a stronger part and the key too" "Um… I'm gonna say not met and the reason for that is taking the graphic as a whole, it's not easy to read. Individual elements are easy to read but as a whole but it's not like I'm	Low

Category	Guideline	Concept	Example Comments	Difficulty
			looking at this for a couple of seconds - but it's taking me time to sit here and look at it and that's maybe because it's trying to get too much info into one graphic but overall, I don't think an easy take away, zero.	

[a] Although interviews were not transcribed I referred to audio-recordings for the inclusion of quotes and example comments included in Table 3.3 are direct quotes from interview participants

[b] The sample graph, randomly selected from the dataset, was black and white, which led to some confusion applying rating scales for color.

CHAPTER 4

RESULTS

My primary research questions for this study were whether the use and quality of data visualizations in reports increased the likelihood that the reports were used, addressed as two separate questions. First, I explored the degree to which the use of data visualizations in reports, regardless of their quality, was related to the frequency reports were used. Second, I explored the degree to which the quality of data visualizations in reports was related to the frequency reports were used. I report the results organized by each research question below. In addition, due to the exploratory nature of my study, I conclude the chapter with a review of relationships identified between alternative predictors of use—the length of reports, the type of report, and user affiliation with a university—and the frequency reports were used.

Did Use of Data Visualizations Increase Use of Reports?

Prior to conducting a regression analysis with frequency of use as the dependent variable, I tested if the data resembled a Poisson, or count, distribution due to the low number of reports in the sample referenced more than once (21%). I recoded the raw frequency of use data, where 1 = 0, 2 = 1, 3 = 2…7 = 6, conducted a one-sample Kolmogorov-Smirnov test for a Poisson distribution and found that the data were consistent with that distribution (Kolmogorov-Smirnov $Z = 0.94$, $p = 0.34$). As such, I ran a Poisson regression with frequency of use as the dependent variable to address the research question. A Poisson regression analysis is appropriate for rare occurrences and can be used to predict an event rate, in this case the rate a report would be used more than once (Azen & Walker, 2011).

I used the percent of data visualizations in reports as the primary predictor variable to answer the question if use of data visualizations in reports was related to the frequency the reports were used, with type of report, user affiliation, and report length as alternative predictors. Descriptive statistics are presented in Table 4.1.

Table 4.1

Descriptive Statistics for Use of Data Visualizations

						Skewness		Kurtosis	
Variable	*N*	Min	Max	Mean	*SD*	Stat	*SE*	Stat	*SE*
Use[a]	215	0	6.00	0.33	0.82	3.88	0.17	18.74	0.33
Visualizations[b]	215	0	1.25	0.10	0.18	2.70	0.17	9.58	0.33
Length[c]	215	2	411.00	53.54	57.98	2.61	0.17	9.10	0.33
Type[d]	215	0	16.00	8.08	5.32	-0.20	0.17	-1.28	0.33
Affiliation[e]	215	0	1.00	0.30	0.46	0.89	0.17	-1.22	0.33

[a] Total number of times a report was referenced.
[b] Percent of data visualizations by total number of pages including appendices.
[c] Total number of pages, including appendices.
[d] Reports were rated on eight criteria where *0* = more like traditional research; *2* = more like advocacy research, and *1* = a mixture of both for a total possible score of 16 points.
[e] Dichotomous variable where *0* = witness was not affiliated with a university and *1* = at least one witness was affiliated with a university.

The descriptive statistics show that multiple variables follow non-normal distributions. Use has strong a strong positive skew (3.88), and both frequency of use and user affiliation have strong kurtosis values (18.74; -1.22), indicating non-normal distributions. This is expected for categorical variables. In addition, the percent of data visualizations and report length have strong positive skews (skew = 2.70 and 2.61 respectively) as well as strong kurtosis values (kurtosis > 9.0), also indicating non-normal distributions. The peaked and skewed shape of report length was

due to most reports (68%) being 50 pages or less and an outlier report that was 411 pages.[10] The peaked and positively skewed shape of the percent of data visualizations is because a majority of the reports in the sample (122, 57%) did not have any data visualizations. Although the covariates do not follow normal distributions, a Poisson regression generalized linear model does not assume normal distribution of either the outcome or predictor variables (Azen & Walker, 2011).

Relationships among Covariates

One of the assumptions of Poisson regression is that there is no collinearity between the independent variables included in the model. To check this assumption, I examined a correlation matrix for multicollinearity among the continuous variables and found a significant relationship between the percent of data visualizations and the type of report, as seen in Table 4.2. I then ran a secondary analysis using linear regression in order to assess the variance inflation factor (VIF) of the independent variables[11] and found good tolerance (0.98 to 0.99) and associated VIF (VIF = 1.01 to 1.03).[12] Thus, although there was a significant correlation among one pair of predictor variables, the VIF values indicated the overlap was not large enough to be a problem in the regression analysis.

[10] All reports included in the sample which were 300 pages or more and organized into chapters were classified as books and removed from the analysis. Report 575 was 411 pages but was not organized into chapters and thus did not meet the criteria for exclusion.
[11] Conducting a linear regression to explore multicollinearity is appropriate even with a categorical outcome as the dependent variable is not considered in the analysis and does not impact relationships among the independent variables (Heck, personal communication, February 23, 2018).
[12] VIF values less than 10 are considered acceptable.

Table 4.2

Correlation Matrix (N = 215)

	Percent of data visualization	Report length	Type of report
Percent of data visualization	1	0.07	0.14*
Report length		1	0.08
Type of report			1

* Correlation is significant at the 0.05 level (2-tailed).

Data Visualizations and Frequency of Use

Prior to conducting the regression analysis, I standardized the continuous predictor variables—percent of data visualization, type of report, and length of report—into *Z*-scores to center the means at 0, causing the intercept to be a report with an average percent of visualizations, average length, average type score, and no affiliation with a university, to allow for more meaningful interpretation of the data.

Table 4.3

Parameter Estimates for the Full Sample

				95% Wald CI		Hypothesis Test			
Parameter	*N*	*B*	*SE*	Lower	Upper	Wald Chi-Square	*df*	*p*	*Exp(B)*
(Intercept)		-1.04	0.22	-1.47	-0.61	22.66	1	0	0.35
Visualizations	215	0.11	0.10	-0.08	0.31	1.27	1	0.26	1.12
Length	215	0.00	0.12	-0.23	0.22	0.00	1	0.99	1.00
Type	215	0.25	0.13	-0.01	0.50	3.57	1	0.06	1.28
Affiliation = 0	215	-0.17	0.27	-0.70	0.36	0.38	1	0.54	0.85
Affiliation = 1		0							1
(Scale)		1[a]							

a. Fixed at the displayed value.

The first research question driving the study was if the use of data visualizations in reports made a difference in the use of the report. For a report of average length, report type, and no affiliation with a university, I did not find a significant relationship between the percent of data visualizations and the frequency reports were used at the 95% confidence level ($p = 0.17$). Moreover, the predicted increase in the event rate (1.12) per each increase in the standard deviation of the percent of data visualizations was about the same as chance.

Does the Quality of Data Visualizations Increase Use of Reports?

The majority of reports (57%) in the full analytic sample ($N = 215$) did not include any data visualizations. Because of this, I explored if the quality of data visualizations was related to the frequency with which reports were used on the sub-sample of 93 reports that had data visualizations in them. Descriptive statistics for the sub-sample are presented in Table 4.4. Based on the distribution of the full sample, I began with a one-sample Kolmogorov-Smirnoff test and confirmed the data followed a Poisson distribution (Kolmogorov-Smirnov $Z = 0.53$, $p = 0.95$). As such, I conducted a second Poisson regression analysis to explore the relationship between the quality of data visualizations and report use.

The descriptive statistics, including high skew and kurtosis values which indicate non-normal distributions, were similar to those in the full sample. However, as mentioned above, a Poisson regression does not assume normal distribution of predictor variables.

I used the Data Visualization Checklist (DVC) percent score as the primary predictor of the frequency reports were used. As a reminder, this was an average across all data visualizations in each report and was calculated based on the total points awarded divided by the total points possible, minus items that were scored as not applicable. The regression analysis included DVC

as the primary predictor of use, with the length of the report, type of report, and the academic affiliation of the user as alternative predictors. Prior to running the regression analyses, I checked for multicollinearity between the continuous predictors—DVC, report length, and report type—using a correlation matrix and confirmed they were independent of each other.

Table 4.4

Descriptive Statistics for Quality of Data Visualizations

						Skewness		Kurtosis	
	N	Min	Max	Mean	*SD*	Stat	*SE*	Stat	*SE*
Use	93	0	5.00	0.34	0.81	3.61	0.25	15.65	0.50
DVC[a]	93	0.56	0.95	0.79	0.07	-0.61	0.25	1.13	0.50
Length	93	4	411.00	66.83	69.93	2.35	0.25	6.77	0.50
Type	93	0	16.00	8.33	5.11	-0.25	0.25	-1.17	0.50
Affiliation	93	0	1.00	0.26	0.44	1.12	0.25	-0.75	0.50

[a] Data Visualization Checklist

Quality of Data Visualizations and Frequency of Use

As with the first regression analysis, I used standardized continuous predictor variables (DVC, report length, and report type) to center the means at 0, causing the intercept to be a report with an average DVC score, average length, and average type to allow for more meaningful interpretation of the data.

A commonly used hypothesis test for categorical variables is Wald's chi-squared test. However, with small sample sizes Wald's chi-squared often has an inflated standard error and underestimates relationships in the data. For this reason log-likelihood ratio chi-squared estimates are preferred with small sample sizes when values for the two estimates differ

(Bewick, Cheek, & Ball, 2005). Because the estimated Wald's and log-likelihood chi-squared values differed, I report the log-likelihood chi-squared values in Table 4.5.

Table 4.5

Parameter Estimates for the Sub-Sample of Reports with Data Visualizations

				Hypothesis Test			
Parameter	*N*	*B*	*SE*	Log-likelihood Ratio Chi-squared	*df*	*p*	*Exp(B)*
(Intercept)	93	-0.94	0.34	45.02	1	0.00	0.39
DVC[a]	93	0.10	0.18	0.31	1	0.58	1.11
Length	93	-0.51	0.28	4.37	1	0.04	0.60
Type	93	0.3	0.20	3.10	1	0.05	1.48
Affiliation	93	-0.38	0.41	0.80	1	0.38	0.67
[affiliation = 1]		0[b]					1

[a] Data Visualization Checklist
[b] Fixed at the displayed value.

The primary predictor of interest in the study was the quality of data visualizations as measured by the DVC. For a report of average length and type score, and the user not affiliated with a university, there was a slight positive association between scores on the DVC and the frequency a report was used. However, the association was not significant at the 95% confidence level ($p = 0.58$), and the predicted increase in the event rate (1.11) per each increase in the standard deviation of the DVC score was about the same as chance.

Alternative Predictors of the Frequency of Use

In addition to answering the research questions if the use and quality of data visualizations were related to the frequency reports were used, I explored if the length of reports,

the type of report, or a user's affiliation with a university predicted the frequency reports were used. In examining the results from the Poisson regression analysis using the full sample (N = 215, see Table 4.3), I found a near-significant relationship between the type of report and frequency reports were used at the 95% confidence level (p = 0.06). This suggests a possible relationship between the two variables that might be significant at the 95% level with a larger sample or fewer predictors. For a report with an average percent of data visualizations, average report type score, and no user affiliation with a university, for each increase in one standard deviation in the report type score we can predict the event rate will increase by a factor of 1.3 (see Table 4.3).

I also explored if the length of reports, type of report, or user affiliation predicted the frequency reports were used for the sub-sample of reports with data visualizations (N = 93, see Table 4.5). For reports with data visualizations, I found a significant negative relationship between the length of reports and the frequency reports were used (p = 0.04). For reports with data visualizations that had an average DVC score, average report type score, and no user affiliation with a university, an increase of one standard deviation in report length predicted the event rate will decrease by a factor of 0.62. In other words, for each increase in the standard deviation in the report length, we can predict a 38% decrease in the rate the report will be used more than once.

For reports with data visualizations (N = 93), I also found a significant, positive association between report type and frequency of use (p = 0.05). For a report with an average DVC score, average length, and no user affiliation with a university, one point increase in the standard deviation of the report type score is predicted to increase the rate a report will be used

more than once by a factor of 1.4. In other words, a higher report type score—indicating a report is more like advocacy research—the greater rate a report will be used more than once.

CHAPTER 5

DISCUSSION

The purpose of this study was to better understand if the inclusion and quality of data visualizations in reports is related to use of the reports, specifically the symbolic use of findings to persuade others. This chapter is organized by different factors potentially related to use identified in the rating study: the use and quality of data visualizations, the length of reports, and the type of reports. In each section, I discuss my results in relation to prior research on evaluation use and/or evaluation communication, as well as new understandings that emerged in the course of the study. I end the chapter by discussing limitations of the study, largely due to the nature of the data used for the analyses.

To my knowledge, this is the first study to empirically examine the relationship between data visualization and evaluation use. I drew from Newman, Brown and Braskamp's (1980) simulation studies on evaluation communication, as well as Evergreen's (2016) unpublished work on the design of reports. However, my study was exploratory. As such, addressing my research questions required additional steps including collecting evidence of validity, training raters, and executing a series of rating exercises to generate data about reports to help explain use of the findings. Due to the complexity and variety of factors that may influence the use of an evaluation, I had low expectations of finding a clear link between data visualizations and use.

Data Visualization and Use

My research questions were if the use and quality of data visualizations were related to the frequency reports were used. Within the realm of congressional testimony on teacher quality, I did not find a significant relationship between the inclusion of data visualizations and the

frequency reports were used. This was also true for the quality of data visualizations and the frequency reports were used. For both, an increase in one standard deviation in either the percent of data visualizations or the Data Visualization Checklist (DVC) score was predicted to increase the likelihood of use by a factor of 1.1, almost the same as chance.

My findings support Evergreen's (2016) preliminary study, which also found that the overall design of research and evaluation reports was not significantly related to the symbolic use of reports.

Report Length and Use

Based on the rationale that policymakers have limited time to read lengthy reports, I included the length of report as an alternative predictor of use, following the idea that busy individuals are less likely to read long reports, and therefore less likely to use them. When I included report length as a predictor on the full sample (N = 215), I did not find a significant relationship between the lengths of reports and the frequency they were used. However, I did find a significant negative relationship between report length and the use of reports with data visualizations (N = 93). Specifically, for reports with an average report type score, each increase in the standard deviation of report length predicted a 38% decrease in the probability of the report being used more than once.

Although prior research on evaluation communication did not consider report length as a factor of use. Evergreen (2011b) examined the length of evaluation reports and found they were, on average, 175 pages. The reports in both the full sample and the sub-sample of just those reports with data visualizations were shorter in length than Evergreen found and this was likely because the sample was not restricted to only evaluation reports.

Important to note when interpreting this result, I did not find significant relationships between the number of data visualizations in reports and report length (N = 215), nor between the quality of data visualizations and report length (N = 93). This suggests there could be something unique about reports with data visualizations in relation to their length that resulted in a decreased likelihood those reports are used more than once. I compared the two samples and found that the sub-sample of reports with data visualizations had a higher average length (67 pages) than the full sample (54 pages). While more work is needed to better understand why report length was only a factor of use for reports with data visualizations, the results do support the theory that shorter reports may promote use.

Type of Report and Use

Drawing from prior research on evaluation use and evaluation communication, I included the type of report as an alternative predictor of use, presented as a measure of the trustworthiness, or credibility, of a report. As a reminder, reports were classified as either more like traditional research or more like advocacy research based on eight criterion including but not limited to who produced it, the tone, and the production quality of the report. For each criteria a report scored a *0* if it modeled traditional research, a *2* if it modeled advocacy research, and a *1* if it was a mixture of both. Following this, reports with higher scores were more like advocacy research and, as such, considered less credible as defined by prior research on evaluation (Alkin & King, Cousins & Leithwood, 1986; Leviton & Hughes, 1981).

I found that for both the full sample (N = 215) and the sub-sample of just those reports with data visualizations (N = 93), the report type score was a significant predictor of use. Holding other predictors constant, for each increase in one standard deviation in the report type score, reports were 1.3 times more likely to be used more than once. For those reports with data

visualizations, an increase in one standard deviation in the report type score predicts reports are 1.4 times more likely to be used more than once.

Higher report type scores indicate a report is more like advocacy research, which has characteristics that should make the report seem less trustworthy. For example, advocacy research is characterized as offering policy recommendations based on anecdotal evidence and lacking an objective tone or reference to other literature (see Table 5.1). Because of this, we would expect that reports with high report type scores would be considered less credible, and therefore would be *less* likely to be used. This expectation was supported by prior research which found credibility, related to objectivity, believability, and use of appropriate methods, was a positive factor in evaluation use (Alkin & King, Cousins & Leithwood, 1986; Leviton & Hughes, 1981). My findings suggest the opposite is true, and reports with characteristics that would make them less credible were used more frequently.

Table 5.1

Advocacy research and policy research as ideal types

Element	**Traditional Research**	**Advocacy Research**
Producer	University based researchers	Think tank or intermediary staff
Recommendations and Evidence	Policy recommendations not mentioned or implied as implications Data analysis or rigorous case studies	Policy recommendations highlighted Uses anecdotal evidence
Style and Production	Tone of objectivity Several citations Standard research paper format	Tone of persuasion Relatively few citations High level production quality

Note: Reproduced from Reckhow et al. (2015, p.8)

One reason reports similar to advocacy research were more likely to be used may be due to the context in which the reports were used. In Shulha and Cousins' (1997) review, they synthesized views put forth by Weiss (1988) during the Weiss/Patton debates, and argued

information often competes for credibility in complex settings. In other words, in certain decision making settings—Congress for example—there are competing priorities and complexities which may color an individual's perception of what is considered credible. In these situations credibility, like beauty, may be in the eye of the beholder.

Another possible reason for the finding may be due to my interpretation of type of report solely as a proxy for credibility, or rather, lack of credibility. It is possible that the coding schema also represents a construct akin to persuasion. A report which has high report type scores is polished, includes recommendations front and center, and the evidence to support them is more likely to be stories or anecdotes than rigorous research. Holistically, the information is presented in a way designed to persuade readers about a particular finding or position.

Although investigating the underlying constructs within the coding schema for type of report is beyond the bounds of the present study, Reckhow et al. (2015) found evidence of more than one construct in the schema. The authors conducted a factor analysis on 106 reports for the eight criteria used to classify the reports as more like traditional, or more like advocacy research. They found all eight criteria loaded onto a common factor, with correlation coefficients equal to or greater than 0.6. However two criteria, report producer and use of citations, also loaded onto a second factor and explained 14% of the overall variance. The authors stopped short of naming or exploring the second factor, but their findings suggest there could be more than one underlying construct within the type of report.

Data Visualizations and Type of Report

One aspect of the coding scheme for type of report included production quality, where more polished reports, described as "magazine" quality by Reckhow and Tompkins-Stange

(2015, p. 32), were coded higher. Important to this study, one of the characteristics used to score the production quality of reports as more like traditional research was if graphs in the report required interpretation. In contrast, research on cognition, graphic design, and the interpretation of graphs promote graphs that simplify interpretation (Evergreen & Metzner, 2013; Ware, 2008, 2012) and include a clear take-away message (Evergreen & Metzner, 2013; Ellis & Dix, 2015). In addition, interpretability was one of the key underlying concepts identified in the DVC. Because of this, the ability to interpret graphs appears to be a component of both the DVC and the type of report score.

Evidence of overlap. Due to the large number of reports without data visualizations in my analytic sample, I only addressed the question if data visualization quality was related to use on a sub-sample of reports which had data visualizations. However, during initial analyses and reviewing descriptive statistics on the full sample, I found a significant positive correlation between report type scores and DVC scores. This relationship went away when I only looked at the sub-sample of reports with data visualizations and therefore was not included in the results. The likely reason the relationship disappeared with the smaller sample was because reports with no data visualizations had DVC scores of zero and, similarly, reports considered traditional research had report type scores equal to zero. The overlap between DVC scores and report type requires more investigation.

Limitations

The primary limitation of the study was using frequency of use as the dependent variable. Due to the nature of the report data used in the study, all of the reports had been used. For this reason I was unable to investigate characteristics that may have led to their initial use and which may be different than the characteristics that predict frequency of use. For this reason, it is

important to be clear that the relationships identified between the length of reports and report type only predict the rates that reports will be used more than once, it does not predict use in general. Extending from this, new ideas about the length and type of report as predictors of use can only be generalized in contexts where reports are used more than once.

In spite of use being very difficult to track, I was able to use existing data from publically available congressional testimony, which provided a clear, well-defined use variable (Reckhow et al., 2015). The benefit of this over prior research on evaluation communication which simulated use (Newman, Brown, & Braskamp, 1980) was the difference between asking if someone would use findings and demonstrating that they did. However, the trade-off to have a clean use variable was a significant limitation to the study's findings due to the political context in which the reports were used. Although I accounted for users' affiliation with a university, I did not have information about a users' political affiliation or if the reports referenced were intended to support or contest specific legislation which might have been controversial or divisive. The complexity of a political landscape introduces potential additional influences which may impact use beyond observable characteristics of a report which were accounted for in the study design. It is possible that the relationships between report type and frequency of use would wash away in the presence of a more important factor of use in the political landscape. At a minimum, this limitation restricts generalizability of the findings to reports referenced within congressional testimony.

An additional limitation was collapsing research and evaluation reports together rather than only looking at evaluation reports to investigate evaluation use. I did find some evidence that the reports in my study differed from related research on evaluation use. For example, Evergreen (2011b) found the average length of evaluation reports sampled from the Information

Education Science track of the National Science Foundation was 175 pages. In my sample the average report length was 58 pages. However, one of the first reviews of research on evaluation use conducted by Leviton and Hughes (1980) drew from evaluation and social science research and the factors they identified as important to use were also found in reviews of only evaluation research (Alkin & King, 2017).

A final limitation, and one similar to the challenge of working within a political context, is there are a great many factors that contribute to use. I only investigated a small sample. I selected my covariates based prior research on evaluation use, as well as Evergreen's (2016) preliminary study and factors identified by Newman, Brown, and Braskamp (1981) in their simulation studies looking at communication theory in relation to use. However, due to the lack of research in this area, it is possible that there are other factors not included in my study that better predict use of research and evaluation reports.

Conclusion

The problem statement driving this study was that evaluation reports often follow the conventions of academic publishing which do not align with what we understand about how humans take in information on a page. This disconnect may be particularly poignant for policymakers who are short of time and may not be trained in how to make sense of academic research. One way to address the disconnect is through the design of data visualizations which tend to catch the eye and are better remembered than text, though with the caveat that complex visualizations are not helpful. As such, the study explored the relationship between the use and design of data visualizations and symbolic use of reports in congressional testimony.

Although I did not find a relationship between the use or quality of data visualizations in reports and the frequency reports were used, I did find the length of reports with data visualizations, and the type of report contributed to the symbolic use of reports in a political context. However, there were a number of significant limitations to these findings due to the context of the study.

Implications for Future Research

In spite of the limitations, the results of this study suggest there is a relationship between the type of report and evaluation use, as well as the length of reports with data visualizations, and evaluation use—with the big caveat that these characteristics predict the rate reports are used more than once and not use in general. Because of this caveat, there remains a need for research on symbolic use where the outcome measure is initial use rather than repeated use.

The relationships between the type of report and frequency of use also raises additional questions, including, what is it about reports similar to advocacy research that promotes symbolic use, and is this true for all reports or only those referenced in congressional testimony? In particular, there is promise in exploring characteristics of persuasion and to what extent these characteristics are related to use, including instrumental and conceptual use. Important to note, exploring properties of persuasion is not a new idea. In their introduction situating their research on communication theory and evaluation use, Newman, Brown, and Braskcamp (1980) argued, "…even when the evaluator limits the report to portrayal or to an exposition of the issues, there is an element of persuasion involved" (p. 30).

In addition, more empirical research is needed on the use of data visualizations in evaluation reporting. Prior work on cognition paints a solid case for the ability of images to aid

in readers' noticing and making sense of complex information (Ware, 2008, 2012). However, this is with the strong caveat that poorly designed data visualizations could hurt interpretation of findings, or worse, lead to misinterpretation due to limitations of working memory (Ellis & Dix, 2015). Although I did not find that data visualizations aid in the symbolic use of findings—at least within a political context—data visualizations were a component of the criteria used to categorize reports as advocacy research, which was found to impact use. For this reason, there is a need to continue exploration of potential relationships between data visualizations and use of evaluation findings, or at a minimum, the interpretation of findings.

Last, the broader idea that the use of design in research or evaluation reports is indicative of non-research, or weak research—for example the "production quality" criteria in Reckhow et al.'s (2015) study—needs further investigation. Although members of the American Evaluation Association are frequently the subject of research on evaluation, it would be helpful to know to what extent evaluators view data visualizations, in particular the concept of sharing a takeaway message, as helpful; and vice versa to what extent do evaluators view the concept as indicative of a lack of credibility.

As a field we are poking at the long-standing bubble that insulates and protects the concept that reports which look like peer-reviewed articles are more credible than those which incorporate design. However, much more research is needed to understand the happy medium between credibility and interpretability.

APPENDIX

Pre-interview Questions

1. Have you seen the checklist before?
2. Have you used it?
3. On a scale from 1-5 how familiar are you with data visualization, where 1 is not at all and 5 is you could teach it?
4. Have you created data visualizations before?
5. If yes, how many years have you been working with them?

Cognitive Interview Steps

1. The interviews were held at locations convenient for interviewees, including but not limited to conference rooms in the offices at or near to the interviewees place of work.
2. I described the purpose of the study to explore the relationship between data visualizations and use of reports and that the Data Visualization Checklist (DVC) would be used to measure the quality of data visualizations.
3. I asked permission to record the interview as a way to take notes.
4. I read aloud and recorded responses to five questions (see Pre-Interview Questions above) about interviewees' exposure to the checklist and experience with data visualizations.
5. I asked if they were ready to get started and explained I would read a script get us started to be sure I did not miss anything important.
6. I read the Interview Script provided below.
7. After interviewees finished responding to the practice exercise prompt, I would ask if they understood what to do.
8. After they confirmed they understood, I gave the Data Visualization Example provided below to the interviewees and reminded them I was going to read out statements on the DVC and they would talk about what they were thinking to get to their rating of 0 = not met, 1 = partially met, 2 = fully met, or "not applicable".
9. I read through the statements on the DVC and asked appropriate follow-up questions depending on their responses (see Think Aloud Conditional Questions below).
10. For the last four statements on the DVC I asked additional probing questions (see Table A1) of all interviewees.
11. I thanked the interviewees for their time, asked if they wanted to receive a copy of the results, and gave them a $10 Starbucks gift card for their time.

Cognitive Interview Script

"Thank you for coming in, I appreciate your time. The purpose of our exercise today is to gather information, not about you, but about the Data Visualization Checklist, to learn more

about how people like yourself think about the different statements in the checklist and use them to rate data visualizations. One of the ways we'll use this information is to improve training for people who will use the checklist to rate data visualizations. I will read you the statements in the checklist and I'd like you to refer to the graph in front of you to come up with a rating. For each statement you can give a rating of 0 = not met, 1 = partially met, 2 = fully met, or "not applicable". For example, the statement "gridlines are muted" is not applicable for a pie graph because it doesn't have gridlines. However, I'd also like to hear about what you're thinking. Please try to think out loud, just tell me everything that comes to mind whether it seems important or not. I may ask questions about how you came to your rating more about your understanding of the statement and I'll be taking lots of notes. If any statement seems unclear, is hard to answer, or doesn't make sense, please tell me. I didn't create the checklist so it won't bother me! We'll just take our time and get as far as we can in an hour. If we have time, I will ask you more about specific statements in the checklist. Do you have any questions? Okay, before we get started let's do a quick practice exercise: Try to visualize the place where you live and think about how many windows are in that place. As you count up the windows, tell me what you are seeing and thinking about (Adapted from a training prompt developed by David Mingay as reported by Willis, 2005)

Think Aloud Conditional Questions[13]

- *Difficulty answering*. What was going through your head as you tried to rate that statement?
- *Delayed response*. You took a little while to rate that statement. What were you thinking about?
- *Uncertainty*. You seem a little unsure. If so, can you tell me why?
- *Error* – response implies misunderstanding. Restate response in a question, "So that graph does not have tick marks?"
- *Request for more information*. If I weren't available or able to answer, what would you decide it means?

Probing Questions

I asked specific probing questions of all interviewees to address anticipated problems with complex and ambiguous statements. An overview of the statement, anticipated problems or issues, and the specific probe are outlined in Table A1.

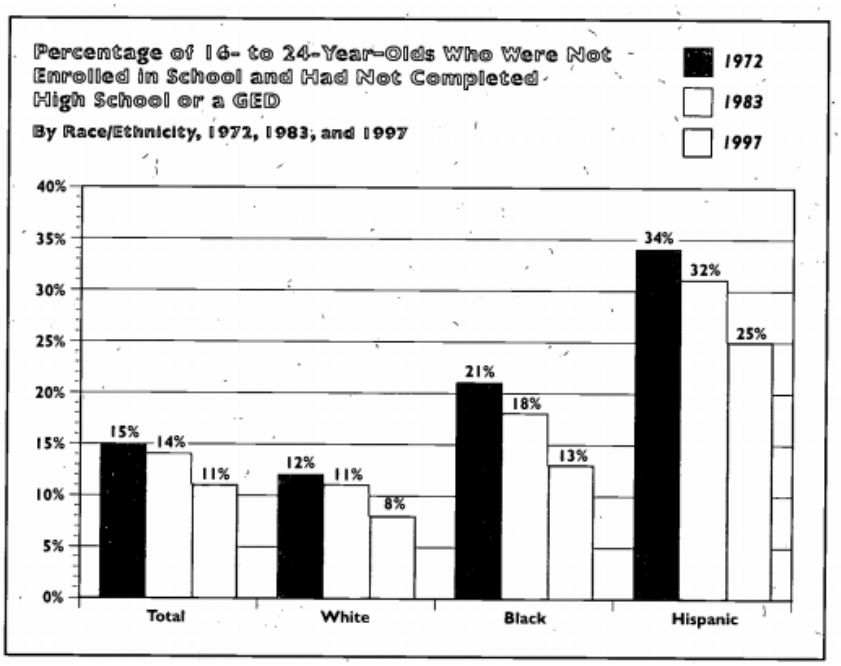
Table A.1

Statement-Specific Probing Questions

[13] Adapted from Willis, 2005

Statement	Issue	Probes
Graph highlights significant finding or conclusion.	Unclear how this is achieved and may cause uncertainty.	How sure are you of your rating? How hard was this to rate?
The type of graph is appropriate for the data.	Requires technical knowledge about graph types.	Can you restate the statement in your own words? What, to you, does "appropriate graph" mean?
Graph has appropriate level of precision.	Requires technical knowledge about graphs. Unclear how this is achieved and may cause uncertainty.	Can you restate the statement in your own words? What, to you, does "appropriate level of precision" mean? How sure are you of your rating?
Individual chart statements work together to reinforce the overarching takeaway message	Unclear how this is achieved and may cause uncertainty.	How sure are you of your rating? How hard was this to rate?

Data Visualization Example for the Think Aloud



Data Visualization Checklist

by Stephanie Evergreen & Ann K. Emery
May 2016

This checklist is meant to be used as a guide for the development of high impact data visualizations. Rate each aspect of the data visualization by circling the most appropriate number, where 2 points means the guideline was fully met, 1 means it was partially met, and 0 means it was not met at all. n/a should not be used frequently, but reserved for when the guideline truly does not apply. For example, a pie chart has no axes lines or tick marks to rate. If the guidelines has been broken intentionally to make a point, rate it n/a and deduct those points from the total possible. Refer to the Data Visualization Anatomy Chart on the last page for guidance on vocabulary and the Resources at the end for more details.

	Guideline	Rating
Text Graphs don't contain much text, so existing text must encapsulate your message and pack a punch.	**6-12 word descriptive title is left-justified in upper left corner** Short titles enable readers to comprehend takeaway messages even while quickly skimming the graph. Rather than a generic phrase, use a descriptive sentence that encapsulates the graph's finding or "so what?" Western cultures start reading in the upper left, so locate the title there.	2 1 0 n/a
	Subtitle and/or annotations provide additional information Subtitles and annotations (call-out text within the graph) can add explanatory and interpretive power to a graph. Use them to answer questions a viewer might have or to highlight specific data points.	2 1 0 n/a
	Text size is hierarchical and readable Titles are in a larger size than subtitles or annotations, which are larger than labels, which are larger than axis labels, which are larger than source information. The smallest text - axis labels - are at least 9 point font size on paper, at least 20 on screen.	2 1 0 n/a
	Text is horizontal Titles, subtitles, annotations, and data labels are horizontal (not vertical or diagonal). Line labels and axis labels can deviate from this rule and still receive full points. Consider switching graph orientation (e.g., from column to bar chart) to make text horizontal.	2 1 0 n/a
	Data are labeled directly Position data labels near the data rather than in a separate legend (e.g., on top of or next to bars and next to lines). Eliminate/embed legends when possible because eye movement back and forth between the legend and the data can interrupt the brain's attempts to interpret the graph.	2 1 0 n/a
	Labels are used sparingly Focus attention by removing the redundancy. For example, in line charts, label every other year on an axis. Do not add numeric labels *and* use a y-axis scale, since this is redundant.	2 1 0 n/a

Arrangement

Improper arrangement of graph elements can confuse readers at best and mislead viewer at worst. Thoughtful arrangement makes a data visualization easier for a viewer to interpret.

Proportions are accurate 2 1 0 n/a
A viewer should be able measure the length or area of the graph with a ruler and find that it matches the relationship in the underlying data. Y-axis scales should be appropriate. Bar charts start axes at 0. Other graphs can have a minimum and maximum scale that reflects what should be an accurate interpretation of the data (e.g., the stock market ticker should not start at 0 or we won't see a meaningful pattern).

Data are intentionally ordered 2 1 0 n/a
Data should be displayed in an order that makes logical sense to the viewer. Data may be ordered by frequency counts (e.g., from greatest to least for nominal categories), by groupings or bins (e.g., histograms), by time period (e.g., line charts), alphabetically, etc. Use an order that supports interpretation of the data.

Axis intervals are equidistant 2 1 0 n/a
The spaces between axis intervals should be the same unit, even if every axis interval isn't labeled. Irregular data collection periods can be noted with markers on a line graph, for example.

Graph is two-dimensional 2 1 0 n/a
Avoid three-dimensional displays, bevels, and other distortions.

Display is free from decoration 2 1 0 n/a
Graph is free from clipart or other illustrations used solely for decoration. Some graphics, like icons, can support interpretation.

Color

Keep culture-laden color connotations in mind. For example, pink is highly associated with feminine qualities in the USA.

Use sites like Color Brewer to find color schemes suitable for reprinting in black-and-white and for colorblindness.

Color scheme is intentional 2 1 0 n/a
Colors should represent brand or other intentional choice, not default color schemes. Use your organization's colors or your client's colors. Work with online tools to identify brand colors and others that are compatible.

Color is used to highlight key patterns 2 1 0 n/a
Action colors should guide the viewer to key parts of the display. Less important, supporting, or comparison data should be a muted color, like gray.

Color is legible when printed in black and white 2 1 0 n/a
When printed or photocopied in black and white, the viewer should still be able to see patterns in the data.

Color is legible for people with colorblindness 2 1 0 n/a
Avoid red-green and yellow-blue combinations when those colors touch one another.

Text sufficiently contrasts background 2 1 0 n/a
Black/very dark text against a white/transparent background is easiest to read.

Lines

Excessive lines—gridlines, borders, tick marks, and axes—can add clutter or noise to a graph, so eliminate them whenever they aren't useful for interpreting the data.

Gridlines, if present, are muted 2 1 0 n/a
Color should be faint gray, not black. Full points if no gridlines are used. Gridlines, even muted, should not be used when the graph includes numeric labels on each data point.

Graph does not have border line 2 1 0 n/a
Graph should bleed into the surrounding page or slide rather than being contained by a border.

Axes do not have unnecessary tick marks or axis lines 2 1 0 n/a
Tick marks can be useful in line graphs (to demarcate each point in time along the y-axis) but are unnecessary in most other graph types. Remove axes lines whenever possible.

Graph has one horizontal and one vertical axis 2 1 0 n/a
Viewers can best interpret one x- and one y-axis. Don't add a second y-axis. Try a connected scatter plot or two graphs, side by side, instead. (A secondary axis used to hack new graph types is ok, so long as viewers aren't being asked to interpret a second y-axis.)

Overall

Graphs will catch a viewer's attention so only visualize the data that needs attention. Too many graphics of unimportant information dilute the power of visualization.

Graph highlights significant finding or conclusion 2 1 0 n/a
Graphs should have a "so what?" – either a practical or statistical significance (or both) to warrant their presence. For example, contextualized or comparison data help the viewer understand the significance of the data and give the graph more interpretive power.

The type of graph is appropriate for data 2 1 0 n/a
Data are displayed using a graph type appropriate for the relationship within the data. For example, change over time is displayed as a line graph, area chart, slope graph, or dot plot.

Graph has appropriate level of precision 2 1 0 n/a
Use a level of precision that meets your audiences' needs. Few numeric labels need decimal places, unless you are speaking with academic peers. Charts intended for public consumption rarely need *p* values listed.

Individual chart elements work together to reinforce the overarching takeaway message 2 1 0 n/a
Choices about graph type, text, arrangement, color, and lines should reinforce the same takeaway message.

For more support, check out:

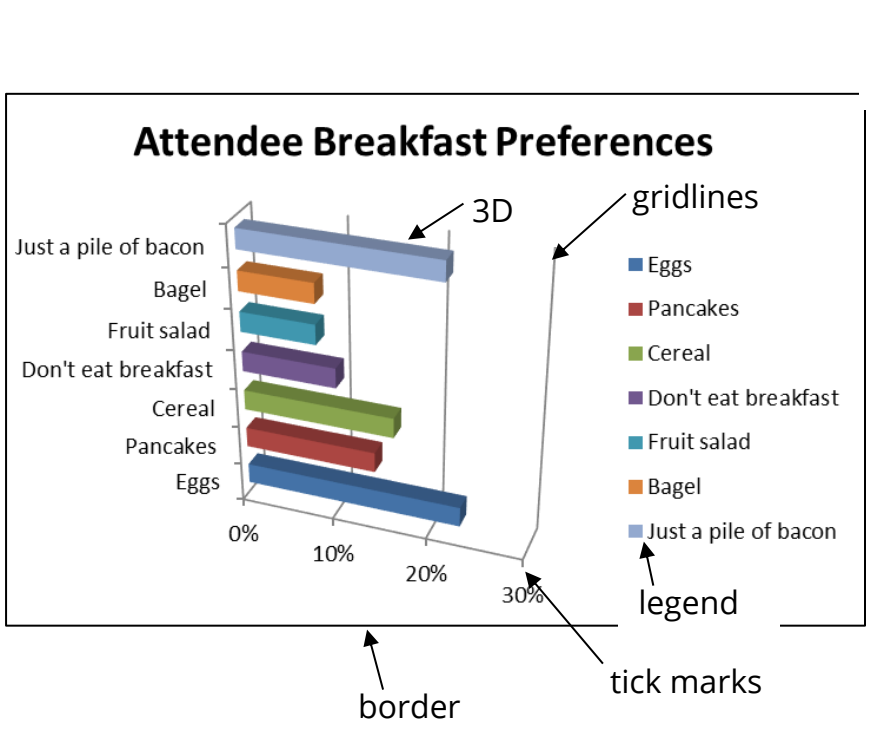
AnnKEmery.com/blog
StephanieEvergreen.com/blog
Stephanie Evergreen's books, *Presenting Data Effectively* & *Effective Data Visualization*

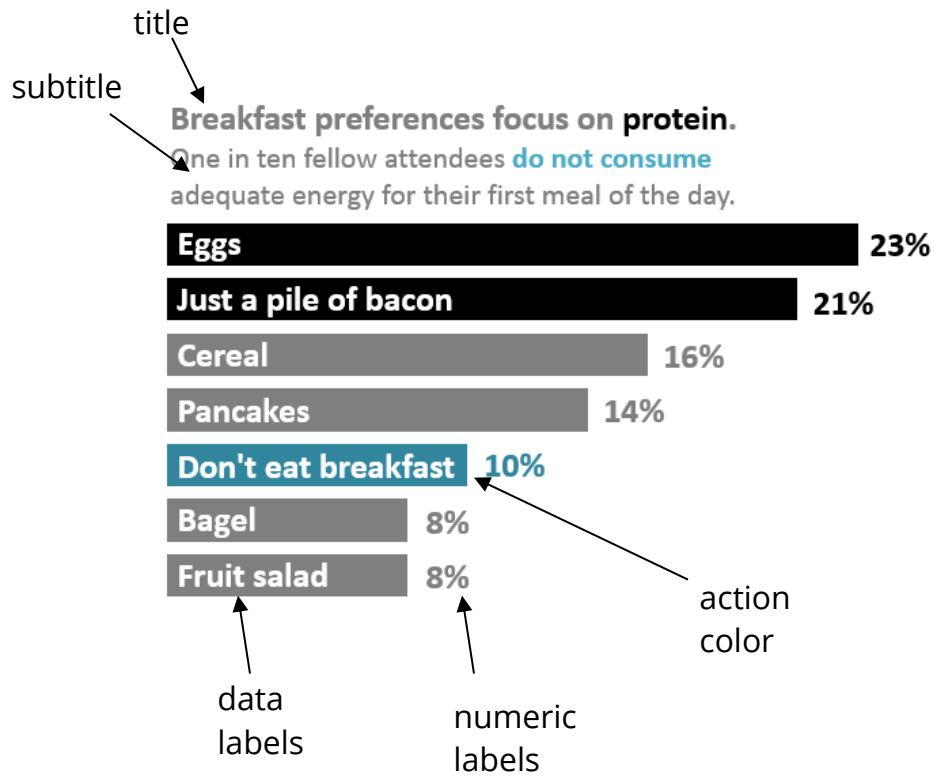
Score: _____ / _____ = _____ %

Well-formatted data visualizations score between 90-100% of available points. At this level, viewers are better able to read, interpret, and retain content.

Data Visualization Anatomy Chart

Confused by the terminology? Review the anatomy charts below for illustration of what's what.





Report Type Coding Schema

Original Coding Schema[14]

Traditional Research. Code "*1*" If…

1) Who produced the research?
    a. University based researchers
    b. Individuals with training in research/methods
2) Conclusions
    a. Discuss caveats
    b. Explain sources of uncertainty
3) Policy recommendations
    a. Not mentioned
    b. Mentioned as potential implications (in conclusion), with caveats
4) Production quality
    a. Research paper style
    b. No color
    c. Few bullet points, charts, or other embellishments; graphs require interpretation
5) Case study
    a. Informative/dispassionate explanation
6) Citations/references
    a. Several
    b. Includes references to scholarly articles/books
7) Tone of objectivity
8) Explanation of Methods
    a. Included in body of paper

Advocacy Research. Code "2" if…

1) Who produced the research?
    a. Staff at think tank or advocacy organization
    b. Individuals without specific training in research/methods
2) Conclusions
    a. Few/no caveats discussed
    b. Little to no discussion of uncertainty
3) Policy recommendations
    a. Specifically highlighted
    b. Mentioned in the introduction
    c. Described without caveats
    d. Mentions current policy issue under consideration by lawmakers

[14] Drawn from Reckhow et al. (2015)

4) Production quality
    a. Color
    b. Glossy or magazine style
    c. Photos, charts, bullet points, and other embellishments
5) Anecdotes
    a. Described as proof/evidence of effectiveness
    b. Inspirational tone
6) Few/lack of citations and references
7) Tone of persuasion
8) Little explanation of methods
    a. Methods presented very briefly or only in the appendix

Expanded Coding Schema

The additional descriptive language was added to the coding schema adopted from Reckhow et al. to clarify criteria used to code reports more or less like traditional or advocacy research. In their study description the authors stated that reports received scores of "1" if they had characteristics of both traditional and advocacy research. We referred to the original codes awarded to five reports to develop guidelines for when to code reports as "0", "1", or "2."

Table A.2

Expanded Coding Schema

Criteria	Description	Scoring	Notes
Producer	Who produced the research	0 = Majority of authors (i.e. 2 of 3) are affiliated with a university/trained in research methods. 2 = Authorship is listed as organization staff or individuals w/out university affiliation. 1 = Partial points if equally co-authored by university affiliated individual or one trained in research methods and organization staff. See Reports 29 and 549 for examples.	If cannot confirm affiliation coded as 2.
Conclusions	Inclusion or not of caveats	0 = Discusses caveats and sources of uncertainty along with conclusions. 2 = Does not discuss or dismisses caveats or sources of uncertainty. 1 = No explicit mention of caveats/uncertainty but includes clear description of parameters, i.e. this finding is based solely on this population, etc.	If uses language such as "must", "should", etc. in conclusion, same as not including caveats or sources of uncertainty and gets a 2.

Criteria	Description	Scoring	Notes
Policy Recommendations	Inclusion and discussion of policy recommendations	0 = No recommendations are included and if they are included they are discussed with caveats/limitations. 2 = Recommendations are highlighted or included in the introduction of a report 1 = Recommendations are included in an executive summary or prior to discussion of findings or are substantiated in the body of the report.	If there are recommendations in the main body of the report it gets a 2. If recommendations are not prominent and supported, gets a 1.
Production Quality	Quality and design of the report	0 = Mirrors traditional research article. May have color use in headings and graphs but if graphs are included, they require interpretation, i.e. the title does not include the takeaway message whereas Report 4 does not. 2 = Use of color and design and/or includes graphs/call-out text that share takeaway findings. Graph titles help the reader interpret the graph. 1 = This is usually a 0 or 2. See differences between Report 4 (coded 0) and Report 29 (coded 2), i.e. Report 29 includes call-outs and graph titles with the takeaway message.	Excessive use of bullet points is a type of graphic design and example of advocacy report and gets a 2.
Evidence	Description/ discussion of study or evidence	0 = Evidence for findings is primary data or a case study discussed in detail. If findings are based on other's work, the work is described in detail in an objective voice - i.e. descriptive with no opinions or commentary. 2 = There is no evidence for findings or the evidence is based on other literature or studies which are not fully described, i.e. So and so found X...without explanation for how they found X. See Report 29 for example. 1 = Something in-between 0 and 2. Example: Evidence for findings is included and fully described but discussed subjectively, i.e. the author's opinion on the evidence is provided.	If can answer what is the data/evidence for the findings/ recommendations and who did you get it from, gets a 0. If the evidence is presented subjectively, i.e. lots of adjectives and adverbs, gets a 1.

Criteria	Description	Scoring	Notes
Citations	Inclusion or not of citations or references	0 = Multiple scholarly citations/references (i.e. journals & books) are included in a notes, endnotes, or reference section, or included in footnotes throughout the paper. 2 = Report does not include citations or references to scholarly works or includes less than five. 1 = Report includes more than five but less than 10 citations or references to scholarly work.	Reference to a scholarly work, must be a journal article, aka has a volume/issue reference or book. If cannot tell, then consider it a non-academic work.
Tone	Objective or persuasive tone	0 = Discussion of study, findings, conclusion, etc. are descriptive and limited to provided evidence. 2 = Discussion of study, findings, conclusion, recommendations, etc. extrapolate beyond provided evidence and/or include words like "Need to", "Must", "Should", etc. 1 = If authors do not extrapolate but do use subjective language when describing evidence. Example, "masterpiece", etc.	
Methods	Inclusion and placement of methods section	0 = Methods are fully discussed in main body of the report 2 = Methods are not discussed or are included in the appendix of a report. 1 = Methods are briefly touched upon. See Report #29 for an example.	If not empirical research and there is no methods section, score as a 2.

REFERENCES

Ali, N., & Peebles, D. (2013). The effect of Gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors, 55*(1), 183–203. http://doi.org/10.1177/0018720812452592

Alkin, M. C., Daillak, R., & White, P. (1979). Using evaluations—Does evaluation make a difference? Thousand Oaks, CA: Sage.

Alkin, M. C., & King, J. A. (2016). The historical development of evaluation use. *American Journal of Evaluation, 37*(4), 568–579. doi.org/10.1177/1098214016665164

Alkin, M. C., & King, J. A. (2017). Definitions of evaluation use and misuse, evaluation influence, and factors affecting use. *American Journal of Evaluation, 38*(3), 434–450. http://doi.org/10.1177/1098214017717015

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Evaluation Association. (2017). *Conference Program*. Retrieved from http://www.evaluationconference.org/p/cm/ld/fid=505

Azen, R. & Walker, C. M. (2011). *Categorical data analysis for the behavioral and social sciences.* New York, NY: Routledge.

Azzam, T., Evergreen, S., Germuth, A. A., & Kistler, S. J. (2013). Data visualization and evaluation. In T. Azzam & S. Evergreen (Eds.), *Data visualization, part 1: New Directions for Evaluation, 139*, 7–32. http://doi.org/10.1002/ev.20065

Azzam, T., & Jacobson, M. R. (2015). Reflections on the future of research on evaluation. In P. R. Brandon (Ed.), *Research on Evaluation: New Directions for Evaluation, 148*, 103–116. http://doi.org/10.1002/ev.20160

Bagby, R. M., Andrew, G. R., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become dead weight? *American Journal of Psychology, 161*(12), 2163–2177.

Baughman, S., Boyd, H. H., & Franz, N. K. (2012). Non–formal educator use of evaluation results. *Evaluation and Program Planning, 35*(3), 329–336. http://doi.org/10.1016/j.evalprogplan.2011.11.008

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care, 9*(1). 112–118.

Brandon, P. R., & Singh, J. M. (2009). The strength of the methodological warrants for the findings of research on program evaluation use. *American Journal of Evaluation, 30*(2), 123–157. http://doi.org/10.1177/1098214009334507

Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: developing best practices based on science and experience. *Quality of Life Research, 18*(9), 1263–1278. http://doi.org/10.1007/s11136-009-9540-9

Campbell, R., Townsend, S. M., Shaw, J., Karim, N., & Markowitz, J. (2015). Can a workbook work? Examining whether a practitioner evaluation toolkit can promote instrumental use. *Evaluation and Program Planning, 52*, 107–117. http://doi.org/10.1016/j.evalprogplan.2015.04.005

Caracelli, V. J. (2000). Evaluation use at the threshold of the twenty-first century. In V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use, New Directions for Evaluation*, *88*, 99–111. http://doi.org/10.1002/ev.1194

Chen, C., & Yu, Y. U. E. (2000). Empirical studies of information visualization: A meta-analysis. *International Journal of Human-Computer Studies, 53*(5), 851–866. http://doi.org//10.1006/ijhc.2000.0422

Coryn, C. L. S., Wilson, L. N., Westine, C. D., Hobson, K. A., Ozeki, S., Fiekowsky, E. L., Greenman II, G. D., & Schröter, D. C. (2017). A decade of research on evaluation: a systematic review of research on evaluation published between 2005 and 2014. *American Journal of Evaluation*, *38*(3), 329–347. http://doi.org/10.1177/1098214016688556

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331–364. http://doi.org/10.3102/00346543056003331

Cousins, J. B., Svensson, K., Szijarto, B., Pinsent, C., Andrew, C., & Sylvestre, J. (2015). Assessing the practice impact of research on evaluation. In P. R. Brandon (Ed.), *Research on evaluation. New Directions for Evaluation*, 148, 73–88. http://doi.org/10.1002/ev.20158

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. http://doi.org/10.1017/S0140525X01003922

Ealy, J. B. (2016). Visualization of kinetics: Stimulating higher-order thinking via visualization. *Journal of Chemical Education, 93*(2), 394–396. http://doi.org/10.1021/acs.jchemed.5b00215

Ellis, G., & Dix, A. (2015). Decision making under uncertainty in visualization? Paper presented at the IEEE VIS2015, Chicago, IL. Retrieved from http://vda.univie.ac.at/uncertainty2015/submissions/ellis_vdmu.pdf

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: Bradford/MIT Press.

Erwin, K., Bond, M., & Jain, A. (2015). Discovering the language of data: Personal pattern languages and the social construction of meaning from big data. *Interdisciplinary Science Reviews, 40*(1), 44–60. http://doi.org/10.1179/0308018814z.000000000104

Evergreen, S. (2011a). Death by boredom: *The role of visual processing theory in written evaluation communication* (Unpublished doctoral dissertation). Western Michigan University. Kalamazoo, Michigan.

Evergreen, S. (2011b). Eval + comm. In S. Mathison (Ed.), *Really new directions in evaluation: Young evaluators' perspectives. New Directions for Evaluation, 131*, 41–45. http://doi.org/10.1002/ev.376

Evergreen, S. (2016). *The link between graphic design and report use*. Paper presented at the American Evaluation Association Annual Conference, October 26, 2016. Atlanta, GA.

Evergreen, S. & Emery, A. K. (2016). *The data visualization checklist*. Retrieved from http://stephanieevergreen.com/wp-content/uploads/2016/10/DataVizChecklist_May2016.pdf

Evergreen, S. & Emery, A. K. (2014). *The data visualization checklist.* Retrieved from http://stephanieevergreen.com/wp-content/uploads/2014/05/DataVizChecklist_May2014.pdf

Evergreen, S., & Metzner, C. (2013). Design principles for data visualization in evaluation. In T. Azzam & S. Evergreen (Eds.), *Data visualization, part 2. New Directions for Evaluation, 140*, 5–20.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use results from a survey of us American Evaluation Association members. *American Journal of Evaluation, 30*(2), 158–175. http://doi.org/10.1177/1098214008331009

Gulliksen, H. (1950). Intrinsic validity. American Psychologist, 5, 511–517.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: a new hypothesis. *Trends in Cognitive Sciences, 11*(6), 236–242. http://doi.org//10.1016/j.tics.2007.04.001

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34.

Henry, T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293–314.

Herbert, J. L. (2014). Researching evaluation influence: A review of the literature. *Evaluation Review, 38*(5). 388–419. http://doi.org/10.1177/0193841X14547230

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377–410. http://doi.org/10.1177/1098214009341660

King, J. A., & Pechman, E. M. (1984). Pinning a wave to the shore: Conceptualizing evaluation use in school systems. *Educational Evaluation and Policy Analysis, 6(*3), 241–251. http://doi.org/10.3102/01623737006003241

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. In V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use, New Directions for Evaluation, 88,* 5–23. http://doi.org/10.1002/ev.1188

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. http://doi.org/10.1016/j.jcm.2016.02.012

Kosslyn, S. M., Kievit, R. A., Russell, A. G., & Shephard, J. M. (2012). PowerPoint® presentation flaws and failures: A psychological analysis. *Frontiers in Psychology, 3*, 1–22. http://doi.org/10.3389/fpsyg.2012.00230

Landis, J. R., & Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometics*, *33*(1), 159–174.

Leviton, L. C., & Hughes, E. F. X. (1981). Research on the utilization of evaluations. *Evaluation Review, 5(*4), 525–548. http://doi.org/10.1177/0193841X8100500405

Lewis, N. R., Harrison, G. M., Ah Sam, A. F., & Brandon, P. R. (2015). Evaluators' perspectives on research on evaluation. In P. R. Brandon (Ed.), *Research on Evaluation: New Directions for Evaluation*, *148*, 89–102. http://doi.org/10.1002/ev.20159

Martens, P. J. & Roos, N. P. (2005). When health services researchers and policy makers interact: Tales for the tectonic plates. *Health Care Policy, 1*(1), 72–84.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.

Newman, D. L., Brown, R. D., & Braskamp, L. A. (1980). Communication theory and the utilization of evaluation. In L. A. Braskamp & R. D. Brown (Eds.), *Utilization of evaluation information, New Directions for Program Evaluation*, *5*, 29–35. http://doi.org/10.1002/ev.1234

Obe, A. S. (2013). Data visualization and beyond: A multi-disciplinary approach to promote user engagement with official statistics. *Statistical Journal of the International Association for Official Statistics, 29*(3), 173–185. http://doi.org/10.3233/SJI-130783

Pankaj, V., & Emery, A. K. (2016). Data placemats: A Facilitative Technique Designed to Enhance Stakeholder Understanding of Data. In R. S. Fierro, A. Schwartz, & D. H. Smart (Eds.), *Evaluation and facilitation, New Directions for Evaluation*, *149*, 81–93. http://doi.org/10.1002/ev.20181

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text.* (3rd Ed.). Newbury Park, CA: Sage.

Powell, T. (2013). The importance of assessments: How portfolios can impact students' self-efficacy and comprehension in an online graphic design course. Retrieved from https://eric.ed.gov.

Preskill, H., & Caracelli, V. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. *Evaluation Practice, 18*(3), 209–225. http://doi.org/10.1016/s0886-1633(97)90028-3

Radvansky, G. A. & Ashcraft, M. H. (2016). *Cognition, 6th ed*. Boston, MA: Pearson.

Rebora, G., & Turri, M. (2011). Critical factors in the use of evaluation in Italian universities. *Higher Education, 61*(5), 531–544. http://doi.org/10.1007/s10734-010-9347-1

Reckhow, S., Holden, L., & Tompkins-Stange, M. (2015). *Patron of ideas: How advocacy research influences the education policy debate*. Paper presented at the American Political Science Association Annual Meeting, September 2015. San Francisco, CA.

Robertson, K. N., & Wingate L. A. (2017). *Checklist for program evaluation report content.* Kalamazoo, MI: EvaluATE, The Evaluation Center, Western Michigan University. Retrieved from http://www.evaluate.org/resources/checklist-evalrpts/

Roseland, D., Lawrenz, F., & Thao, M. (2015). The relationship between involvement in and use of evaluation in multi-site evaluations. *Evaluation and Program Planning*, *48*, 75–82. http://doi.org/10.1016/j.evalprogplan.2014.10.003

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *Evaluation Practice, 18*(3), 195–208. http://doi.org/10.1016/s0886-1633(97)90027-1

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. http://doi.org/10.1037/0033-2909.86.2.420

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1), 83–117. http://doi.org/10.1023/a:1006985528729

Sjetne, I. S., Iversen, H. H., & Kjøllesdal, J. G. (2015). A questionnaire to measure women's experiences with pregnancy, birth and postnatal care: instrument development and assessment following a national survey in Norway. *BMC Pregnancy & Childbirth, 15*(1), 1–11. http://doi.org/10.1186/s12884-015-0611-3

Skau, D., Harrison, L., & Kosara, R. (2015). An evaluation of the impact of visual embellishments in bar charts. *Computer Graphics Forum, 34*(3), 221–230. http://doi.org/10.1111/cgf.12634

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs,* 74 (11), 1-29. http://doi.org/10.1037/h0093759

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology*, *18*(6), 813–847.

Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage.

Torres, R. T., Preskill, H., & Piontek, M. E. (1996). *Evaluation strategies for communicating and reporting*. Thousand Oaks, CA.

Tomlinson, D., Mol Van Otterloo, S., O'Sullivan, C., Gibson, P., Johnston, D. L., Portwine, C., Spiegler, B., Baggott, C., Tolend, M., Dupuis, L. L., & Sung, L. (2016). Methodological issues identified during cognitive interviews in the development of a pediatric cancer symptom screening tool. *Psycho-Oncology, 25*(3), 349–353. http://doi.org/10.1002/pon.3821

Tsuji, B. H., & Lindgaard, G. (2014). *Comparing novices and experts in their exploration of data in line graphs.* Paper presented at the International Conference on Cognition and Exploratory Learning in Digital Age (CELDA), Porto, Portugal, October 25–27, 2014.

Tufte, E. R. (2006). *The cognitive style of PowerPoint: Pitching out corrupts within* (2nd ed.). Cheshire, CT: Graphics Press.

Turnbull, B. (1999). The mediating effect of participation efficacy on evaluation use. *Evaluation and Program Planning, 22*(2), 131–140. http://doi.org/10.1016/s0149-7189(99)00012-9

University of Hawai‘i at Mānoa. (2014). *Electronic Thesis and Dissertation Style and Policy Guide.* Honolulu, HI: University of Hawai‘i at Mānoa.

Vallin, L. M., Philippoff, J., Pierce, S., & Brandon, P. R. (2015). Research-on-evaluation articles published in the American Journal of Evaluation, 1998–2014. In P. R. Brandon (Ed.), *Research on evaluation, New Directions for Evaluation, 148,* 7–15. http://doi.org/10.1002/ev.20153

Ware, C. (2008). *Visual thinking for design.* Burlington, MA: Morgan Kaufmann Publishers.

Ware, C. (2012). *Information Visualization: Perception for Design*. San Francisco, CA: Elsevier Science.

Weiss, C. H. (1967). *Utilization of evaluation: Toward a comparative study*. Washington, DC: Government Printing Office.

Weiss, C. H. (1972). Utilization of evaluation: Toward comparative study. In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education* (pp. 318–326). Boston, MA: Allyn and Bacon.

Weiss, C. H. (1979). The many meanings of research utilization. *Public Administration Review, 39*(5), 426–431. http://doi.org/10.2307/3109916

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*(1), 21–33. http://doi.org/10.1177/109821409801900103

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Sage Publications: Thousand Oaks, CA.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440*(7080), 91–95. http://www.doi.org/10.1038/nature04262

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.